Journal of Machine Learning Research 15 (2014) 1549-1591 Submitted 2/12; Revised 12/13; Published 4/14

High-Dimensional Covariance Decomposition into Sparse Markov and Independence Models

Majid Janzamin MJANZAMI@UCI.EDU
Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Department of Electrical Engineering and Computer Science
University of California
Irvine, CA 92697, USA

Editor: Bin Yu

Abstract

Fitting high-dimensional data involves a delicate tradeoff between faithful representation and the use of sparse models. Too often, sparsity assumptions on the fitted model are too restrictive to provide a faithful representation of the observed data. In this paper, we present a novel framework incorporating sparsity in different domains. We decompose the observed covariance matrix into a sparse Gaussian Markov model (with a sparse precision matrix) and a sparse independence model (with a sparse covariance matrix). Our framework incorporates sparse covariance and sparse precision estimation as special cases and thus introduces a richer class of high-dimensional models. We characterize sufficient conditions for identifiability of the two models, viz., Markov and independence models. We propose an efficient decomposition method based on a modification of the popular ℓ_1-penalized maximum-likelihood estimator (ℓ_1-MLE). We establish that our estimator is consistent in both the domains, i.e., it successfully recovers the supports of both Markov and independence models, when the number of samples n scales as $n = \Omega(d^2 \log p)$, where p is the number of variables and d is the maximum node degree in the Markov model. Our experiments validate these results and also demonstrate that our models have better inference accuracy under simple algorithms such as loopy belief propagation.

Keywords: high-dimensional covariance estimation, sparse graphical model selection, sparse covariance models, sparsistency, convex optimization

1. Introduction

Covariance estimation is a classical problem in multi-variate statistics. The idea that second-order statistics capture important and relevant relationships between a given set of variables is natural. Finding the sample covariance matrix based on observed data is straightforward and widely used (Anderson, 1984). However, the sample covariance matrix is ill-behaved in high-dimensions, where the number of dimensions p is typically much larger than the number of available samples n ($p \gg n$). Here, the problem of covariance estimation is ill-posed since the number of unknown parameters is larger than the number of available samples, and the sample covariance matrix becomes singular in this regime.

Various solutions have been proposed for high-dimensional covariance estimation. Intuitively, by restricting the class of covariance models to those with a limited number of free parameters, we can successfully estimate the models in high dimensions. A natural mech-

anism to achieve this is to impose a sparsity constraint on the covariance matrix. In other words, it is presumed that there are only a few (off-diagonal) non-zero entries in the covariance matrix, which implies that the variables under consideration approximately satisfy *marginal independence*, corresponding to the zero pattern of the covariance matrix (Kauermann, 1996) (and we refer to such models as independence models). Many works have studied this setting and have provided guarantees for high-dimensional estimation through simple thresholding of the sample covariance matrix and other related schemes. See Section 1.2. In many settings, however, marginal independence is too restrictive and does not hold. For instance, consider the dependence between the monthly stock returns of various companies listed on the S&P 100 index. It is quite possible that a wide range of complex (and unobserved) factors such as the economic climate, interest rates etc., affect the returns of all the companies. Thus, it is not realistic to model the stock returns of various companies through a sparse covariance model.

A popular alternative sparse model, based on *conditional independence* relationships, has gained widespread acceptance in recent years (Lauritzen, 1996). In this case, sparsity is imposed *not* on the covariance matrix, but on the inverse covariance or the *precision* matrix. It can be shown that the zero pattern of the precision matrix corresponds to a set of conditional-independence relationships and such models are referred to as graphical or Markov models. Going back to the stock market example, a first-order approximation is to model the companies in different divisions[1] as conditionally independent given the S&P 100 index variable, which captures the overall trends of the stock returns, and thus removes much of the dependence between the companies in different divisions. High-dimensional estimation in models with sparse precision matrices has been widely studied, and guarantees for estimation have been provided under a set of sufficient conditions. See Section 1.2 for related works. However, sparse Markov models may not be always sufficient to capture all the statistical relationships among variables. Going back to the stock market example, the approximation of using the S&P index node to capture the dependence between companies of different divisions may not be enough. For instance, there can still be a large *residual* dependence between the companies in manufacturing and mining divisions, which cannot be accounted by the S&P index node.

In this paper, we consider decomposition of the observed data into two domains, viz., Markov and independence domains. We posit that the observed data results in a sparse graphical model under structured perturbations in the form of an independence model, see Figure 1. This framework encapsulates Markov and independence models, and incorporates a richer class of models which can faithfully capture complex relationships, such as in the stock market example above, and yet retain parsimonious representation. The idea that a combination of Markov and independence models can provide good model-fitting is not by itself new and perhaps the work which is closest to ours is the work by Choi et al. (2010), where multi-resolution models with a known hierarchy of variables is considered. Their model consists of a combination of a sparse precision matrix, which captures the conditional independence across scales, and a sparse covariance matrix, which captures the residual in-scale correlations. Heuristics for learning and inference are provided in Choi et al. (2010). However, the approach in Choi et al. (2010) has several deficiencies, including

1. See `http://www.osha.gov/pls/imis/sic_manual.html` for classifications of the companies.

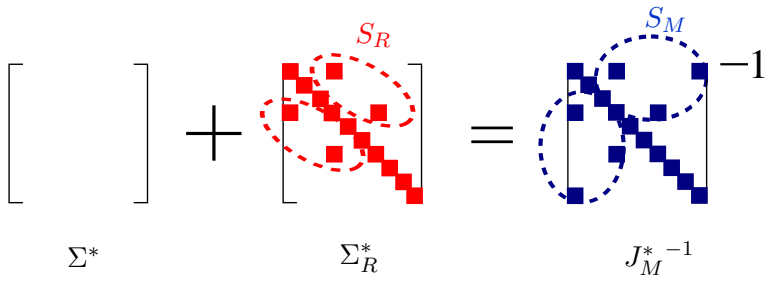


Figure 1: Representation of the covariance decomposition problem, where perturbing the observed covariance matrix with a structured noise model results in a sparse graphical model. The case where the noise model has sparse marginal dependencies is considered.

lack of theoretical guarantees, assumption of a known sparsity support for the Markov model, use of expectation maximization (EM) which has no guarantees of reaching the global optimum, non-identifiability due to the presence of both latent variables and residual correlations, and so on. In contrast, we develop efficient convex optimization methods for decomposition, which are easily implementable and also provide theoretical guarantees for successful recovery. In summary, in this paper, we provide an in-depth study of efficient methods and guarantees for joint estimation of a combination of Markov and independence models.

Our model reduces to sparse covariance and sparse inverse covariance estimation for certain choices of tuning parameter. Therefore, we incorporate a range of models from sparse covariance to sparse inverse covariance.

1.1 Summary of Contributions

We consider joint estimation of Markov and independence models, given observed data in a high dimensional setting. Our contributions in this paper are three fold. First, we derive a set of sufficient restrictions, under which there is a unique decomposition into the two domains, viz., the Markov and the independence domains, thereby leading to an *identifiable* model. Second, we propose novel and efficient estimators for obtaining the decomposition, under both exact and sample statistics. Third, we provide strong theoretical guarantees for high-dimensional learning, both in terms of norm guarantees and *sparsistency* in each domain, viz., the Markov and the independence domain.

Our learning method is based on convex optimization. We adapt the popular ℓ_1-penalized maximum likelihood estimator (MLE), proposed originally for sparse Markov model selection and has efficient implementation in the form of graphical lasso (Friedman et al., 2007). This method involves an ℓ_1 penalty on the precision matrix, which is a convex relaxation of the ℓ_0 penalty, in order to encourage sparsity in the precision matrix. The Lagrangian dual of this program is a *maximum entropy* solution which approximately fits the given sample covariance matrix. We modify this program to our setting as follows: we incorporate an additional ℓ_1 penalty term involving the residual covariance matrix (corresponding to the independence model) in the max-entropy program. This term can be viewed as encouraging sparsity in the independence domain, while fitting a maximum en-

tropy Markov model to the rest of the sample correlations. We characterize the optimal solution of the above program, and also provide intuitions on the class of Markov and independence model combinations which can be incorporated under this framework. As a byproduct of this analysis, we obtain a set of conditions for identifiability of the two model components.

We provide strong theoretical guarantees for our proposed method under a set of sufficient conditions. We establish that it is possible to obtain *sparsistency* and norm guarantees in both the Markov and the independence domains. We establish that the number of samples n is required to scale as $n = \Omega(d^2 \log p)$ for consistency, where p is the number of variables, and d is the maximum degree in the Markov graph. The set of sufficient conditions for successful recovery are based on the so-called notion of *mutual incoherence*, which controls the dependence between different sets of variables (Ravikumar et al., 2011). In Section 7, the synthetic experiments are run on a model which does not necessarily satisfy sufficient mutual incoherence conditions; But we observe that our method has good numerical estimation performance even when the above incoherence conditions are not fully satisfied.

We establish that our estimation reduces to sparse covariance and sparse inverse covariance estimation for certain choices of tuning parameter. On one end, it reduces to the ℓ_1 penalized MLE for sparse precision estimation (Ravikumar et al., 2011). On the other extreme, it reduces to (soft) threshold estimator for sparse covariance estimator, on lines of Bickel and Levina (2008). Moreover, our conditions for successful recovery are similar to those previously characterized for consistent estimation of sparse covariance/precision matrix.

Our experiments validate our theoretical results on the sample complexity and demonstrate that our method is able to learn a richer class of models, compared to sparse graphical model selection, while requiring similar number of samples. In particular, our method is able to provide better estimates for the overall precision matrix, which is dense in general, while the performance of ℓ_1-based optimization is worse since it attempts to approximate the dense matrix via a sparse estimate. Additionally, we demonstrate that our estimated models have better accuracy under simple distributed inference algorithms such as loopy belief propagation (LBP). This is because the Markov components of the estimated models tend to be more *walk summable* (Malioutov et al., 2006), since some of the correlations can be "transferred" to the residual matrix. Thus, in addition to learning a richer model class, incorporating sparsity in both covariance and precision domains, we also learn models amenable to efficient inference. We also apply our method to real data sets. We see the resulting models are fairly interpretable for the real data sets. For instance, for stock returns data set, we observe in both Markov and residual graphs that there exist edges among companies in the same division or industry, e.g., in the residual graph, nodes "HD", "WMT", "TGT" and "MCD", all belonging to division Retail Trade form a partition. Also for foreign exchange rate data set, we observe that the statistical dependencies of foreign exchange rates are correlated with the geographical locations of countries, e.g., it is observed in the learned model that the exchange rates of Asian countries are more correlated.

1.2 Related Works

There have been numerous works on high-dimensional covariance selection and estimation, and we describe them below. In all the settings below based on sparsity of the covariance matrix in some basis, the notion of consistent estimation of the sparse support is known as *sparsistency*.

Sparse Graphical Models: Estimation of covariance matrices by exploiting the sparsity pattern in the inverse covariance or the precision matrix has a long history. The sparsity pattern of the precision matrix corresponds to a Markov graph of a graphical model which characterizes the set of conditional independence relationships between the variables. Chow and Liu established that the maximum likelihood estimate (MLE) for tree graphical models reduces to a maximum weighted spanning tree algorithm where the edge weights correspond to empirical mutual information. The seminal work by Dempster (1972) on covariance selection over chordal graphs analyzed the convex program corresponding to the Gaussian MLE and its dual, when the graph structure is known.

In the high-dimensional regime, penalized likelihood methods have been used in a number of works to achieve parsimony in covariance selection. Penalized MLE based on ℓ_1 penalty has been used in Huang et al. (2006); Meinshausen and Bühlmann (2006); d'Aspremont et al. (2008); Banerjee et al. (2008); Rothman et al. (2008); Ravikumar et al. (2011), among numerous other works, where sparsistency and norm guarantees for recovery in high dimensions are provided. Graphical lasso (Friedman et al., 2007) is an efficient and popular implementation for the ℓ_1-MLE. There have also been recent extensions to group sparsity structures(Yuan and Lin, 2006; Zhao et al., 2009), scenarios with missing samples (Loh and Wainwright, 2011) , semi-parametric settings based on non-paranormals (Liu et al., 2009), and to the non-parametric setting (Kolar et al., 2010). In addition to the convex methods, there have also been a number of non-convex methods for Gaussian graphical model selection (Spirtes and Meek, 1995; Kalisch and Bühlmann, 2007; Zhang, 2009; Anandkumar et al., 2011; Zhang, 2008). While we base much of our consistency analysis on Ravikumar et al. (2011), we also need to develop novel techniques to handle the delicate issue of errors in the two domains, viz., Markov and independence domains.

Sparse Covariance Matrices: In contrast to the above formulation, alternatively we can impose sparsity on the covariance matrix. Note that the zero pattern in the covariance matrix corresponds to marginal independence relationships (Cox and Wermuth, 1993; Kauermann, 1996; Banerjee and Richardson, 2003). High-dimensional estimation of sparse covariance models has been extensively studied in El Karoui (2008); Bickel and Levina (2008); Cai et al. (2010), among others. Wagaman and Levina (2009) consider block-diagonal and banded covariance matrices and propose an Isomap method for discovering meaningful orderings of variables. The work in Lam and Fan (2009) provides unified results for sparsistency under different sparsity assumptions, viz., sparsity in precision matrices, covariance matrices and models with sparse Cholesky decomposition.

The above works provide strong guarantees for covariance selection and estimation under various sparsity assumptions. However, they cannot handle matrices which are combinations of different sparse representations, but are otherwise dense when restricted to any single representation.

Decomposable Regularizers: Recent works have considered model decomposition based on observed samples into desired parts through convex relaxation approaches. Typically, each part is represented as an *algebraic variety*, which are based on *semi-algebraic* sets, and conditions for recovery of each component are characterized. For instance, decomposition of the inverse covariance matrix into sparse and low-rank varieties is considered in Chandrasekaran et al. (2009, 2010a); Candès et al. (2009) and is relevant for latent Gaussian graphical model. The work in Silva et al. (2011) considers finding a sparse-approximation using a small number of positive semi-definite (PSD) matrices, where the "basis" or the set of PSD matrices is specified a priori. In Negahban et al. (2010), a unified framework is provided for high-dimensional analysis of the so-called M-estimators, which optimize the sum of a convex loss function with decomposable regularizers. A general framework for decomposition into a specified set of algebraic varieties was studied in Chandrasekaran et al. (2010b).

The above formulations, however, cannot incorporate our scenario, which consists of a combination of sparse Markov and independence graphs. This is because, although the constraints on the inverse covariance matrix (Markov graph) and the covariance matrix (independence graph) can each be specified in a straightforward manner, their combined constraints on the resulting covariance matrix is not easy to incorporate into a learning method. In particular, we do not have a decomposable regularizer for this setting.

Multi-Resolution Models: Perhaps the work which is closest to ours is the work by Choi et al. (2010), where multi-resolution models with a known hierarchy of variables is considered. The model consists of a combination of a sparse precision matrix, which captures the conditional independence across scales, and a sparse covariance matrix, which captures the residual in-scale correlations. Heuristics for learning and inference are provided. However, the work has three main deficiencies: the sparsity support is assumed to be known, the proposed heuristics have no theoretical guarantees for success and the models considered are in general not identifiable, due to the presence of both latent variables and residual correlations.

2. Preliminaries and Problem Statement

Notation: For any vector $v \in \mathbb{R}^p$ and a real number $a \in [1, \infty)$, the notation $\|v\|_a$ refers to the ℓ_a norm of vector v given by $\|v\|_a := \left(\sum_{i=1}^p |v_i|^a\right)^{\frac{1}{a}}$. For any matrix $U \in \mathbb{R}^{p \times p}$, the induced or the operator norm is given by $\|\!|U|\!\|_{a,b} := \max_{\|z\|_a=1} \|Uz\|_b$ for parameters $a, b \in [1, \infty)$. Specifically, we use the ℓ_∞ operator norm which is equivalent to $\|\!|U|\!\|_\infty = \max_{i=1,\ldots,p} \sum_{j=1}^p |U_{ij}|$. We also have $\|\!|U|\!\|_1 = \|\!|U^T|\!\|_\infty$. Another induced norm is the spectral norm $\|\!|U|\!\|_2$ (or $\|\!|U|\!\|$) which is equivalent to the maximum singular value of U. We also use the ℓ_∞ element-wise norm notation $\|U\|_\infty$ to refer to the maximum absolute value of the entries of U. Note that it is not a matrix norm but a norm on the vectorized form of the matrix. The trace inner product of two matrices is denoted by $\langle U, V \rangle := \mathrm{Tr}(U^T V) = \sum_{i,j} U_{ij} V_{ij}$. Finally, we use the usual notation for asymptotics: $f(n) = \Omega(g(n))$ if $f(n) \geq cg(n)$ for some constant $c > 0$ and $f(n) = O(g(n))$ if $f(n) \leq c' g(n)$ for some constant $c' < \infty$.

2.1 Gaussian Graphical Models

A Gaussian graphical model is a family of jointly Gaussian distributions which factor in accordance to a given graph. Given a graph $G = (V, E)$, with $V = \{1, \ldots, p\}$, consider a vector of Gaussian random variables $\mathbf{X} = [X_1, X_2, \ldots, X_p]$, where each node $i \in V$ is associated with a scalar Gaussian random variable X_i. A Gaussian graphical model Markov on G has a probability density function (pdf) that may be parameterized as

$$f_{\mathbf{X}}(\mathbf{x}) \propto \exp\left[-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \mathbf{h}^T \mathbf{x}\right], \tag{1}$$

where J is a positive-definite symmetric matrix whose sparsity pattern corresponds to that of the graph G. More precisely,

$$J(i,j) = 0 \iff (i,j) \notin G.$$

The matrix J is known as the potential or concentration matrix, the non-zero entries $J(i,j)$ as the edge potentials, and the vector $\mathbf{h}$ as the potential vector. The form of parameterization in (1) is known as the information form and is related to the standard mean-covariance parameterization of the Gaussian distribution as

$$\boldsymbol{\mu} = J^{-1}\mathbf{h}, \quad \Sigma = J^{-1},$$

where $\boldsymbol{\mu} := \mathbb{E}[\mathbf{X}]$ is the mean vector and $\Sigma := \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$ is the covariance matrix.

We say that a jointly Gaussian random vector $\mathbf{X}$ with joint pdf $f(\mathbf{x})$ satisfies local Markov property with respect to a graph G if

$$f(x_i | \mathbf{x}_{\mathcal{N}(i)}) = f(x_i | \mathbf{x}_{V \setminus i})$$

holds for all nodes $i \in V$, where $\mathcal{N}(i)$ denotes the set of neighbors of node $i \in V$ and, $V \setminus i$ denotes the set of all nodes excluding i. More generally, we say that $\mathbf{X}$ satisfies the global Markov property, if for all disjoint sets $A, B \subset V$, we have

$$f(\mathbf{x}_A, \mathbf{x}_B | \mathbf{x}_S) = f(\mathbf{x}_A | \mathbf{x}_S) f(\mathbf{x}_B | \mathbf{x}_S).$$

where set S is a *separator*[2] of A and B. The local and global Markov properties are equivalent for non-degenerate Gaussian distributions (Lauritzen, 1996).

On lines of the above description of graphical models, consider the class of Gaussian models[3] $\mathcal{N}(\mu, \Sigma_{G_c})$, where the covariance matrix is supported on a graph G_c (henceforth referred to as the conjugate graph), i.e.,

$$\Sigma_{G_c}(i,j) = 0 \equiv (i,j) \notin G_c.$$

Recall that uncorrelated Gaussian variables are independent, and thus,

$$X_i \perp\!\!\!\perp X_j \equiv (i,j) \notin G_c.$$

2. A set $S \subset V$ is a separator for sets A and B if the removal of nodes in S partitions A and B into distinct components.
3. In the sequel, we denote the Markov graph, corresponding the support of the information matrix, as G and the conjugate graph, corresponding to the support of the covariance matrix, as G_c.

Equivalence between pairwise independence and global Markov properties were studied in Cox and Wermuth (1993); Kauermann (1996); Banerjee and Richardson (2003).

In this paper, we posit that the observed model results in a sparse graphical model under structure perturbations in the form of an independence model:

$$\Sigma^* + \Sigma_R^* = {J_M^*}^{-1}, \quad \text{Supp}(J_M^*) = G_M, \text{Supp}(\Sigma_R^*) = G_R, \tag{2}$$

where $\text{Supp}(\cdot)$ denotes the set of non-zero (off-diagonal) entries, G_M denotes the Markov graph and G_R, the independence graph.

2.2 Problem Statement

We now give a detailed description of our problem statement, which consists of the covariance decomposition problem (given exact statistics) and covariance estimation problem (given a set of samples).

2.2.1 Covariance Decomposition Problem

A fundamental question to be addressed is the identifiability of the model parameters.

Definition 1 (Identifiability) *A parametric model $\{P_\theta : \theta \in \Theta\}$ is identifiable with respect to a measure μ if there do not exist two distinct parameters $\theta_1 \neq \theta_2$ such that $P_{\theta_1} = P_{\theta_2}$ almost everywhere with respect to μ.*

Thus, if a model is not identifiable, there is no hope of estimating the model parameters from observed data. A Gaussian graphical model (with no hidden variables) belongs to the family of standard exponential distributions (Wainwright and Jordan, 2008, Ch. 3). Under non-degeneracy conditions, it is also in the minimal form, and as such is identifiable (Brown, 1986). In our setting in (2), however, identifiability is not straightforward to address, and forms an important component of the covariance decomposition problem, described below.

Decomposition Problem: Given the covariance matrix $\Sigma^* = {J_M^*}^{-1} - \Sigma_R^*$ as in (2), where J_M^* is an unknown concentration matrix and Σ_R^* is an unknown residual covariance matrix, how and under what conditions can we uniquely recover J_M^* and Σ_R^* from Σ^*?

In other words, we want to address whether the matrices J_M^* and Σ_R^* are *identifiable*, given Σ^*, and if so, how can we design efficient methods to recover them. If we do not impose any additional restrictions, there exists an *equivalence class* of models which form solutions to the decomposition problem. For instance, we can model Σ^* entirely through an independence model ($\Sigma^* = \Sigma_R^*$), or through a Markov model ($\Sigma^* = {J_M^*}^{-1}$). However, in most scenarios, these extreme cases are not desirable, since they result in dense models, while we are interested in sparse representations with a parsimonious use of edges in both the graphs, viz., the Markov and the independence graphs. In Section 3.1, we provide a sufficient set of structural and parametric conditions to guarantee identifiability of the Markov and the independence components, and in Section 3.2, we propose an optimization program to obtain them.

2.2.2 Covariance Estimation Problem

In the above decomposition problem, we assume that the exact covariance matrix Σ^* is known. However, in practice, we only have access to samples, and we describe this setting below.

Denote $\widehat{\Sigma}^n$ as the sample covariance matrix[4]

$$\widehat{\Sigma}^n := \frac{1}{n}\sum_{k=1}^{n} x_{(k)}x_{(k)}^T, \tag{3}$$

where $x_{(k)}, k = 1, ..., n$ are n i.i.d. observations of a zero mean Gaussian random vector $X \sim \mathcal{N}(0, \Sigma^*)$, where $X := (X_1, ..., X_p)$. Now the estimation problem is described below.

Estimation Problem: Assume that there exists a unique decomposition $\Sigma^* = {J_M^*}^{-1} - \Sigma_R^*$ where J_M^* is an unknown concentration matrix with bounded entries and Σ_R^* is an unknown sparse residual covariance matrix given a set of constraints. Given the sample covariance matrix $\widehat{\Sigma}^n$, our goal is to find estimates of J_M^* and Σ_R^* with provable guarantees.

In the sequel, we relate the exact and the sample versions of the decomposition problem. In Section 4, we propose a modified optimization program to obtain efficient estimates of the Markov and independence components. Under a set of sufficient conditions, we provide guarantees in terms of *sparsistency*, *sign consistency*, and *norm* guarantees, defined below.

Definition 2 (Estimation Guarantees) *We say that an estimate $(\widehat{J}_M, \widehat{\Sigma}_R)$ to the decomposition problem in (2), given a sample covariance matrix $\widehat{\Sigma}^n$, is sparsistent or model consistent, if the supports of $\widehat{J}_M$ and $\widehat{\Sigma}_R$ coincide with the supports of J_M^* and Σ_R^* respectively. It is said to be sign consistent, if additionally, the respective signs coincide. The norm guarantees on the estimates is in terms of bounds on $\|\widehat{J}_M - J_M^*\|$ and $\|\widehat{\Sigma}_R - \Sigma_R^*\|$, under some norm $\|\cdot\|$.*

3. Analysis under Exact Statistics

In this section, we provide the results under exact statistics.

3.1 Conditions for Unique Decomposition

We first provide a set of sufficient conditions under which we can guarantee that the decomposition of Σ^* in (2) into concentration matrix J_M^* and residual matrix Σ_R^* is unique.[5] We impose the following set of constraints on the two matrices:

(A.0) Σ^* and J_M^* are positive definite matrices, i.e., $\Sigma^* \succ 0, J_M^* \succ 0$.

(A.1) Off-diagonal entries of J_M^* are bounded from above, i.e., $\|J_M^*\|_{\infty,\text{off}} \leq \lambda^*$, for some $\lambda^* > 0$.

4. Without loss of generality, we limit our analysis to zero-mean Gaussian models. The results can be easily generalized to models with non-zero means.

5. We drop the positive definite constraint on the residual matrix Σ_R^* thereby allowing for a richer class of covariance decomposition. In Section 5.3, we modify the conditions and the learning method to incorporate positive definite residual matrices Σ_R^*.

(A.2) Diagonal entries of Σ_R^* are zero: $\big(\Sigma_R^*\big)_{ii} = 0$, and the support of its off-diagonal entries satisfies

$$\big(\Sigma_R^*\big)_{ij} \neq 0 \iff |\big(J_M^*\big)_{ij}| = \lambda^*, \quad \forall\, i \neq j.$$

(A.3) For any i, j, we have $\text{sign}\big(\big(\Sigma_R^*\big)_{ij}\big).\text{sign}\big(\big(J_M^*\big)_{ij}\big) \geq 0$, i.e, the signs are the same.

Indeed, the above constraints restrict the class of models for which we can provide guarantees. However, in many scenarios, the above assumptions may be reasonable, and we now provide some justifications. (A.0) is a natural assumption to impose since we are interested in valid Σ^* and J_M^* matrices. Condition (A.1) corresponds to bounded off-diagonal entries of J_M^*. Intuitively, this limits the extent of "dependence" between the variables in the Markov model, and can lead to models where inference can be performed with good accuracy using simple algorithms such as belief propagation. Condition (A.2) limits the support of the residual matrix Σ_R^*: the residual covariances are captured at those locations (edges) where the concentration entries $(J_M^*)_{i,j}$ are "clipped" (i.e., the bound λ^* is achieved). Intuitively, the Markov matrix J_M^* is unable to capture all the correlations between the node pairs due to clipping, and the residual matrix Σ_R^* captures the remaining correlations at the clipped locations. Condition (A.3) additionally characterizes the signs of the entries of Σ_R^*. For the special case, when the Markov model is attractive, i.e., $(J_M^*)_{i,j} \leq 0$ for $i \neq j$, the residual entries $(\Sigma_R^*)_{i,j}$ are also all negative. This implies that the model corresponding to Σ^* is also attractive, since it only consists of positive correlations. By default, we set the diagonal entries of the residual matrix to zero in (A.2) and thus, assume that the Markov matrix captures all the variances in the model. In Section 4.2.1, we provide a simple example of a Markov chain and a residual covariance model satisfying the above conditions.

It is also worth mentioning that the number of model parameters satisfying above conditions is equivalent to the number of parameters in the special case of sparse inverse covariance estimation when $\lambda \to \infty$ (Ravikumar et al., 2011). It is assumed in assumption (A.2) that the residual matrix Σ_R^* takes nonzero value when the corresponding entry in the Markov matrix J_M^* takes its maximum absolute value λ^*. This assumption in conjunction with the sign assumption in (A.3), exactly determines the Markov entry $\big(J_M\big)_{ij}$ when the corresponding residual entry $\big(\Sigma_R\big)_{ij} \neq 0$. So, for each (i, j) pair, only one of the entries $\big(J_M\big)_{ij}$ and $\big(\Sigma_R\big)_{ij}$ are unknown which results that the proposed model in this paper does not introduce additional parameters comparing to the sparse inverse covariance estimation, which is interesting.

According to the above discussion, we observe that the overall covariance and inverse covariance matrices Σ^* and $J^* = \Sigma^{*-1}$ are dense, but represented with small number of parameters. It is interesting that we are able to represent models with dense patterns, but it is important to notice that the sparse representation leads to some restrictions on the model.

In the sequel, we propose an efficient method to recover the respective matrices J_M^* and Σ_R^* under conditions (A.0)-(A.3) and then establish the uniqueness of the decomposition. Finally, note that we do not impose any sparsity constraints on the concentration matrix J_M^*, and in fact, our method and guarantees allow for dense matrices J_M^*, when the exact

covariance matrix Σ^* is available. However, when only samples are available, we limit ourselves to sparse J_M^* and provide learning guarantees in the high-dimensional regime, where the number of samples can be much smaller than the number of variables.

3.2 Formulation of the Optimization Program

We now propose a method based on convex optimization for obtaining (J_M^*, Σ_R^*) given the covariance matrix Σ^* in (2). Consider the following program

$$\begin{aligned}\left(\widehat{\Sigma}_M, \widehat{\Sigma}_R\right) := \underset{\Sigma_M \succ 0, \Sigma_R}{\arg\max} \log\det \Sigma_M - \lambda \|\Sigma_R\|_{1,\text{off}} \\ \text{s.t.} \quad \Sigma_M - \Sigma_R = \Sigma^*,\ (\Sigma_R)_d = 0,\end{aligned} \tag{4}$$

where $\|\cdot\|_{1,\text{off}}$ denotes the ℓ_1 norm of the off-diagonal entries, which is the sum of the absolute values of the off-diagonal entries, and $(\cdot)_d$ denotes the diagonal entries. Intuitively, the parameter λ imposes a penalty on large residual covariances, and under favorable conditions, can encourage sparsity in the residual matrix. The program in (4) can be recast

$$\begin{aligned}\left(\widehat{\Sigma}_M, \widehat{\Sigma}_R\right) := \underset{\Sigma_M \succ 0, \Sigma_R}{\arg\max} \log\det \Sigma_M \\ \text{s.t.} \quad \Sigma_M - \Sigma_R = \Sigma^*,\ (\Sigma_R)_d = 0, \|\Sigma_R\|_{1,\text{off}} \leq C(\lambda),\end{aligned} \tag{5}$$

for some constant $C(\lambda)$ depending on λ. The objective function in the above program corresponds to the entropy of the Markov model (modulo a scaling and a shift factor) (Cover and Thomas, 2006), and thus, intuitively, the above program looks for the optimal Markov model with maximum entropy subject to an ℓ_1 constraint on the residual matrix.

We declare the optimal solution $\widehat{\Sigma}_R$ in (4) as the estimate of the residual matrix Σ_R^*, and $\widehat{J}_M := \widehat{\Sigma}_M^{-1}$ as the estimate of the Markov concentration matrix J_M^*. The justification behind these estimates is based on the fact that the Lagrangian dual of the program in (4) is (see Appendix A)

$$\begin{aligned}\widehat{J}_M := \underset{J_M \succ 0}{\arg\min} \langle \Sigma^*, J_M \rangle - \log\det J_M \\ \text{s.t.} \ \|J_M\|_{\infty,\text{off}} \leq \lambda,\end{aligned} \tag{6}$$

where $\|\cdot\|_{\infty,\text{off}}$ denotes the ℓ_∞ element-wise norm of the off-diagonal entries, which is the maximum absolute value of the off-diagonal entries. Further, we show in Appendix A that the following relations exist between the optimal primal[6] solution $\widehat{J}_M$ and the optimal dual solution $\left(\widehat{\Sigma}_M, \widehat{\Sigma}_R\right)$: $\widehat{J}_M = \widehat{\Sigma}_M^{-1}$, and thus, $\widehat{J}_M^{-1} - \widehat{\Sigma}_R = \Sigma^*$ is a valid decomposition of the covariance matrix Σ^*.

Remark 3 *Notice that when the ℓ_∞ constraint is removed in the primal program in (6), which is equivalent to letting $\lambda \to \infty$, the program corresponds to the maximum likelihood estimate, and the optimal solution in this case is $\widehat{J}_M = {\Sigma^*}^{-1}$. Similarly, in the dual program in (4), when $\lambda \to \infty$, the optimal solution corresponds to $\widehat{\Sigma}_M = \Sigma^*$ and $\widehat{\Sigma}_R = 0$. At the*

6. Henceforth, we refer to the program in (6) as the primal program and the program in (4) as the dual program.

other extreme, when $\lambda \to 0$, $\widehat{J}_M$ is a diagonal matrix, and the residual matrix $\widehat{\Sigma}_R$ is in general, a full matrix (except for the diagonal entries). Thus, the parameter λ allows us to carefully tune the contributions of the Markov and residual components, and we notice in our experiments in Section 7 that λ plays a crucial role in obtaining efficient decomposition into Markov and residual components.

3.3 Guarantees and Main Results

We now establish that the optimal solutions of the proposed optimization programs in (4) and (6) lead to a unique decomposition of the given covariance matrix Σ^* under conditions (A.0)–(A.3) given in Section 3.1.

Theorem 4 (Uniqueness of Decomposition) *Under (A.0)–(A.3), given a covariance matrix Σ^*, if we set the parameter $\lambda = \|J_M^*\|_{\infty,\text{off}}$ in the optimization program in* (4)*, then the optimal solutions of primal-dual optimization programs* (6) *and* (4) *are given by $\big(\widehat{J}_M, \widehat{\Sigma}_R\big) = \big(J_M^*, \Sigma_R^*\big)$, and the decomposition is unique.*

See the proof in Appendix C.

Thus, we establish that the proposed optimization programs in (4) and (6) *uniquely* recover the Markov concentration matrix J_M^* and the residual covariance matrix Σ_R^* given Σ^* under conditions (A.0)–(A.3).

4. Sample Analysis of the Algorithm

In this section, we provide the results under sample statistics where some i.i.d. samples of random variables are only available.

4.1 Optimization Program

We have so far provided guarantees on unique decomposition given the exact covariance matrix Σ^*. We now consider the case, when n i.i.d. samples are available from $\mathcal{N}(0, \Sigma^*)$, which allows us to estimate the sample covariance matrix $\widehat{\Sigma}^n$, as in (3).

We now modify the dual program in (4), considered in the previous section, to incorporate the sample covariance matrix $\widehat{\Sigma}^n$ as follows

$$\begin{aligned} \big(\widehat{\Sigma}_M, \widehat{\Sigma}_R\big) := \underset{\Sigma_M, \Sigma_R}{\arg\max}\ & \log\det \Sigma_M - \lambda \|\Sigma_R\|_{1,\text{off}} \\ \text{s.t. } & \|\widehat{\Sigma}^n - \Sigma_M + \Sigma_R\|_{\infty,\text{off}} \leq \gamma, \\ & \big(\Sigma_M\big)_d = \big(\widehat{\Sigma}^n\big)_d,\ \big(\Sigma_R\big)_d = 0, \\ & \Sigma_M \succ 0, \Sigma_M - \Sigma_R \succ 0. \end{aligned} \tag{7}$$

Note that, in addition to substituting Σ^* by $\widehat{\Sigma}^n$, there are two more modifications in the above program comparing to the exact case in (4). First, the positive-definiteness constraint on the overall covariance matrix $\Sigma = \Sigma_M - \Sigma_R$ is added to make sure that the overall covariance matrix estimation is valid. This constraint is not required in the exact case since we have the constraint $\Sigma = \Sigma^*$ in that case which ensures the positive-definiteness of

overall covariance matrix according to assumption (A.0) that $\Sigma^* \succ 0$. Second, the equality constraint $\Sigma_M - \Sigma_R = \Sigma^*$ is relaxed on the off-diagonal entries by introducing the new parameter γ which allows some deviation. More discussion including the Lagrangian primal form of the above optimization program and the effect of new parameter γ is provided in section 6.

4.2 Assumptions under Sample Statistics

We now provide conditions under which we can provide guarantees for estimating the Markov model J_M^* and the residual model Σ_R^*, given the sample covariance $\widehat{\Sigma}^n$ in high dimensions. These are conditions in addition to conditions (A.0)–(A.3) in Section 3.1.

The additional assumptions for successful recovery in high dimensions are based on the Hessian of the objective function in the optimization program in (19), with respect to the variable J_M, evaluated at the true Markov model J_M^*. The Hessian of this function is given by Boyd and Vandenberghe (2004)

$$\Gamma^* = {J_M^*}^{-1} \otimes {J_M^*}^{-1} = \Sigma_M^* \otimes \Sigma_M^*, \tag{8}$$

where $\otimes$ denotes the Kronecker matrix product (Horn and Johnson, 1985). Thus Γ^* is a $p^2 \times p^2$ matrix indexed by the node pairs. Based on the results for exponential families (Brown, 1986), $\Gamma^*_{(i,j),(k,l)} = \mathrm{Cov}\{X_i X_j, X_k X_l\}$, and hence it can be interpreted as an edge-based alternative to the usual covariance matrix Σ_M^*. Define K_M as the ℓ_∞ operator norm of the covariance matrix of the Markov model

$$K_M := |\!|\!| \Sigma_M^* |\!|\!|_\infty.$$

We now denote the supports of the Markov and residual models. Denote $E_M := \{(i,j) \in V \times V | i \neq j, \left(J_M^*\right)_{ij} \neq 0\}$ as the edge set of Markov matrix J_M^*. Define

$$S_M := E_M \cup \{(i,i) | i = 1, ..., p\}, \tag{9}$$

$$S_R := \{(i,j) \in V \times V | \left(\Sigma_R^*\right)_{ij} \neq 0\}. \tag{10}$$

Thus, the set S_M includes diagonal entries and also all edges of Markov graph corresponding to J_M^*. Also, recall from (A.2) in Section 3.1 that the diagonal entries of Σ_R^* are set to zero, and that the support set S_R is contained in S_M, i.e., $S_R \subset S_M$. Let S_M^c and S_R^c denote the respective complement sets. Define

$$S := S_M \cap S_R^c, \tag{11}$$

so that $\{S_R, S, S_M^c\}$ forms a partition of $\{(1, ..., p) \times (1, ..., p)\}$. This partitioning plays a crucial role in being able to provide learning guarantees. Define the maximum node degree for Markov model J_M^* as

$$d := \max_{j=1,...,p} |\{i : (i,j) \in S_M\}|.$$

Finally, for any two subsets T and T' of $V \times V$, $\Gamma^*_{TT'}$ denotes the submatrix of Γ^* indexed by T as rows and T' as columns. We now impose various constraints on the submatrices of the Hessian in (8), limited to each of the sets $\{S_R, S, S_M^c\}$.

(A.4) **Mutual Incoherence**: These conditions impose mutual incoherence among three partitions of Γ^* indexed by S_R, S_M^c and S. For some $\alpha \in (0,1]$, we have

$$\max\{\|\Gamma^*_{S_M^c S}\big(\Gamma^*_{SS}\big)^{-1}\Gamma^*_{SS_R} - \Gamma^*_{S_M^c S_R}\|_\infty, \|\Gamma^*_{S_M^c S}\big(\Gamma^*_{SS}\big)^{-1}\|_\infty\} \leq (1-\alpha), \tag{12}$$

$$K_{SS_R} := \|\big(\Gamma^*_{SS}\big)^{-1}\Gamma^*_{SS_R}\|_\infty < \frac{1}{4}. \tag{13}$$

(A.5) **Covariance Control**: For the same α specified above, we have the bound:

$$K_{SS} := \|\big(\Gamma^*_{SS}\big)^{-1}\|_\infty \leq \frac{(m-4)\alpha}{4(m-(m-1)\alpha)} \text{ for some } m > 4. \tag{14}$$

(A.6) **Eigenvalue Control**: The minimum eigenvalue of overall covariance matrix Σ^* satisfies the lower bound

$$\lambda_{\min}(\Sigma^*) \geq C_6 d\sqrt{\frac{\log(4p^\tau)}{n}} + C_7 d^2 \frac{\log(4p^\tau)}{n} \text{ for some } C_6, C_7 > 0 \text{ and } \tau > 2.$$

In (A.4), the condition in (12) bounds the effect of the non-edges of the Markov model, indexed by S_M^c, to its edges, indexed by S_R and S. Note that we distinguish between the common edges of the Markov model with the residual model (S_R) and the remaining edges of the Markov model (S). The second condition in (13) controls the influence of the edge-based terms which are shared with the residual matrix, indexed by S_R, to other edges of the Markov model, indexed by $S = S_M \cap S_R^c$. Condition (A.5) imposes ℓ_∞ bounds on the rows of $(\Gamma^*_{SS})^{-1}$. Note that for sufficiently large m, the bound in (14) tends to $\frac{\alpha}{4(1-\alpha)}$. Also note that the conditions (A.4) and (A.5) are only imposed on the Markov model J^*_M and there are no additional constraints on the residual matrix Σ^*_R (other than the conditions previously introduced in Section 3.1). In condition (A.6), it is assumed that the minimum eigenvalue of overall covariance matrix Σ^* is sufficiently far from zero to make sure that its estimation $\widehat{\Sigma}$ is positive definite and therefore a valid covariance matrix.

4.2.1 Example of a Markov Chain + Residual Covariance Model

In this section, we propose a simple model satisfying assumptions (A.0)–(A.5). Consider a Markov chain with concentration matrix J^*_M over 4 nodes, as shown in Figure 2. The diagonal entries in the corresponding covariance matrix $\Sigma^*_M = {J^*_M}^{-1}$ are set to unity, and the correlations between the neighbors in J^*_M are set uniformly to some value $\rho \in (-1,1)$, i.e., $\big(\Sigma^*_M\big)_{ij} = \rho$ for $(i,j) \in E_M$. Due to the Markov property, the correlations between other node pairs are given by $\big(\Sigma^*_M\big)_{13} = \big(\Sigma^*_M\big)_{24} = \rho^2$ and $\big(\Sigma^*_M\big)_{14} = \rho^3$. For the residual covariance matrix Σ^*_R, we consider one edge between nodes 1 and 2, i.e., $S_R = \{(1,2),(2,1)\}$. It is easy to see that conditions (A.0)–(A.2) are satisfied. Recall that $S_M^c = \{(i,j) : (i,j) \notin E_M\}$ and the remaining node pairs belongs to set $S := S_M \setminus S_R$. Through some straightforward calculations, we can show that for any $|\rho| < 0.07$, the mutual incoherence conditions in (A.4) and (A.5) are satisfied for $\alpha = 0.855$ and $m \geq 83$. Note that the value of nonzero entries of Σ^*_R are not involved or restricted by these assumptions. However, they do need to satisfy the sign condition in (A.3). Thus, we have non-trivial models

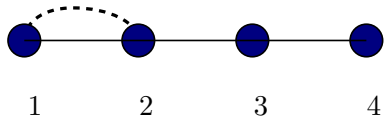


Figure 2: Example of a Markov chain and a residual covariance matrix, where a residual edge is present between nodes 1 and 2.

satisfying the set of sufficient conditions for successful high-dimensional estimation.[7] In Section 7, the synthetic experiments are run on a model which does not necessarily satisfy mutual incoherence conditions (A.4) and (A.5); But we observe that our method has good numerical estimation performance even when the above incoherence conditions are not fully satisfied.

4.3 Guarantees and Main Results

We are now ready to provide the main result of this paper.

Theorem 5 *Consider a Gaussian distribution with covariance matrix* $\Sigma^* = {J_M^*}^{-1} - \Sigma_R^*$ *satisfying conditions (A.0)-(A.6). Given a sample covariance matrix* $\widehat{\Sigma}^n$ *using* n *i.i.d. samples from the Gaussian model, let* $\big(\widehat{J}_M, \widehat{\Sigma}_R\big)$ *denote the optimal solutions of the primal-dual pair* (19) *and* (7)*, with parameters* $\gamma = C_1\sqrt{\log p/n}$ *and* $\lambda = \lambda^* + C_2\sqrt{\log p/n}$ *for some constants* $C_1, C_2 > 0$*, where* $\lambda^* := \|J_M^*\|_{\infty,\text{off}}$*. Suppose that* $\big(\Sigma_R^*\big)_{\min} := \min_{(i,j)\in S_R} \big|\big(\Sigma_R^*\big)_{ij}\big|$ *scales as* $\big(\Sigma_R^*\big)_{\min} = \Omega\big(\sqrt{\log p/n}\big)$ *and the sample size* n *is lower bounded as*

$$n = \Omega\big(d^2 \log p\big), \tag{15}$$

then with probability greater than $1 - 1/p^c \to 1$ *(for some* $c > 0$*), we have:*

a) The estimates $\widehat{J}_M \succ 0$ *and* $\widehat{\Sigma}_R$ *satisfy* ℓ_∞ *bounds*

$$\|\widehat{J}_M - J_M^*\|_\infty = O\bigg(\sqrt{\frac{\log p}{n}}\bigg),$$
$$\|\widehat{\Sigma}_R - \Sigma_R^*\|_\infty = O\bigg(\sqrt{\frac{\log p}{n}}\bigg).$$

b) The estimate $\widehat{\Sigma}_R$ *is sparsistent and sign consistent with* Σ_R^**.*

c) If in addition, $\big(J_M^*\big)_{\min} := \min_{(i,j)\in S_M} \big|\big(J_M^*\big)_{ij}\big|$ *scales as* $\big(J_M^*\big)_{\min} = \Omega\big(\sqrt{\log p/n}\big)$*, then the estimate* $\widehat{J}_M$ *is sparsistent and sign consistent with* J_M^**.*

7. Similarly, for the case when the correlations corresponding to Markov edges are distinct as $\big(\Sigma_M^*\big)_{12} = \rho_1, \big(\Sigma_M^*\big)_{23} = \rho_2$, and $\big(\Sigma_M^*\big)_{34} = \rho_3$, we can argue the same conditions. For compatibility with Figure 2, assume that ρ_1 is the maximum among these three parameters, and therefore, the residual edge is between nodes 1 and 2. This is because the maximum of off-diagonal entries of J_M^* also happens in entry $(1,2)$. Then, the same condition $|\rho_1| < 0.07$ is sufficient for satisfying conditions (A.0)–(A.5).

Proof See Appendix D. ■

Remark 6 *Here, we provide a few more observations and extensions as follows.*

1. **Non-asymptotic sample complexity and error bounds:** *In the above theorem, we establish that the number of samples is required to scale as* $n = \Omega(d^2 \log p)$. *In fact, our results are non-asymptotic, and the exact constants are provided in inequality* (31). *The non-asymptotic form of error bounds are also provided in* (34) *and* (40).

2. **Extension to sub-Gaussian and other distributions:** *In the above theorem, we considered Gaussian distribution. Similar to high dimensional covariance estimation in Ravikumar et al. (2011), the result in the theorem can be easily extended to sub-Gaussian and other distributions with known tail conditions.*

3. **Comparison between direct estimation of Σ^* and the above decomposition:** *The overall matrix* Σ^* *(and* J^**) is a full matrix in general. Thus, if we want to estimate it directly, we need* $n = \Omega\big(p^2 \log p\big)$ *samples since the maximum node degree is* $\Theta(p)$. *Therefore, we can not estimate it directly in high dimensional regime and it demonstrates the importance of such sparse covariance + inverse covariance models for estimation.*

We discussed Remark 3 that the parameter λ allows us to carefully tune the contributions of the Markov and residual components. When $\lambda \to \infty$, the program corresponds to ℓ_1-penalized maximum likelihood estimator which is well-studied in Ravikumar et al. (2011); Rothman et al. (2008). In this case, $\widehat{\Sigma}_R = 0$ and all the dependencies among random variables are captured by the sparse graphical model represented by $\widehat{J}_M$. On the other extreme, when $\lambda^* = 0$ and thus $\lambda = C_2\sqrt{\log p/n} \to 0$, with increasing the number of samples n, the off-diagonal entries in $\widehat{J}_M$ are bounded too tight by λ (refer to the primal program in (19)) and therefore the residual covariance matrix $\widehat{\Sigma}_R$ captures most of the dependencies among random variables. In this case, we have the covariance estimation $\widehat{\Sigma} = \widehat{\Sigma}_M - \widehat{\Sigma}_R$, where the diagonal entries are included in $\widehat{\Sigma}_M$ and the off-diagonal entries are mostly included in $-\widehat{\Sigma}_R$. In order to explain the results for these cases in a more concrete way, we explicitly mention the results for both sparse inverse covariance estimation ($\lambda \to \infty$) and sparse covariance estimation ($\lambda \approx 0$) methods in the following subsections. Note that both of these are special cases of the general result expressed in Theorem 5. Thus, in Theorem 5, we generalize these extreme cases to models with a linear combination of sparse covariance and sparse inverse covariance matrices.

5. Discussions and Extension

In this section, we first provide a detailed discussion of special cases sparse covariance and sparse inverse covariance estimation. Then, the extension of results to the structured noise model is mentioned.

5.1 Sparse Inverse Covariance Estimation

In this section, we mention the result for sparse inverse covariance estimation in high dimensional regime. This result is provided by Ravikumar et al. (2011) and is a special case of Theorem 5 when the parameter λ goes to infinity. Before proposing the explicit result in Corollary 7, we state how the required conditions in Theorem 5 reduces to the conditions in Ravikumar et al. (2011).
Since the support of residual matrix Σ_R^* is a zero matrix in this special case, the mutual incoherence conditions in (A.4) reduce exactly to the same mutual incoherence condition in Ravikumar et al. (2011) as

$$\|\Gamma^*_{S^cS}\big(\Gamma^*_{SS}\big)^{-1}\|_\infty \leq (1-\alpha) \text{ for some } \alpha \in (0,1], \tag{16}$$

where $S = S_M$ is the support of Markov matrix $J^* = J^*_M$ as defined in (9). Also note that the covariance control condition (A.5) is not required any more.
Furthermore, the sample complexity and convergence rate of J^*_M estimation in Theorem 5 exactly reduce to the results in Ravikumar et al. (2011) as (for $q = 8, l = 3$)

$$n > \overline{n}_f\left(p^\tau; 1/\max\left\{v_*, 2ld\Big(1+\frac{q}{\alpha}\Big)K_{SS}K_M \max\Big\{1, \frac{2}{l-1}\Big(1+\frac{q}{\alpha}\Big)K_{SS}K_M^2\Big\}\right\}\right), \tag{17}$$

$$\|\widehat{J} - J^*\|_\infty \leq 2K_{SS}\Big(1+\frac{q}{\alpha}\Big)\overline{\delta}_f(p^\tau; n), \tag{18}$$

where the result is valid for any $q \geq 8$ and $l > 1$.

Corollary 7 (Sparse Inverse Covariance Estimation (Ravikumar et al., 2011)) *Consider a Gaussian distribution with covariance matrix $\Sigma^* = J^{*-1}$ satisfying mutual incoherence condition* (16)*. Given a sample covariance matrix $\widehat{\Sigma}^n$ using n i.i.d. samples from the Gaussian model, let $\widehat{J}$ denote the optimal solution of the primal-dual pair* (19) *and* (7)*, with parameters $\gamma = C_1\sqrt{\log p/n}$ and $\lambda \to \infty$ (removing ℓ_∞ constraints in the primal program* (19)*) for some constant $C_1 > 0$. Suppose that the sample size n is lower bounded as*

$$n = \Omega\big(d^2 \log p\big),$$

then with probability greater than $1 - 1/p^c \to 1$ (for some $c > 0$), we have:

a) The estimate $\widehat{J} \succ 0$ satisfies ℓ_∞ bound

$$\|\widehat{J} - J^*\|_\infty = O\Big(\sqrt{\frac{\log p}{n}}\Big).$$

b) If in addition $\big(J^*\big)_{\min} := \min_{(i,j)\in S_M} |\big(J^*\big)_{ij}|$ scales as $\big(J^*\big)_{\min} = \Omega\big(\sqrt{\log p/n}\big)$, the estimate $\widehat{J}$ is sparsistent and sign consistent with J^*.

Remark 8 (Comparison of general result in Theorem 5 and sparse inverse covariance estimation in Corollary 7) *Considering the results in Theorem 5, sample complexity and convergence rate of estimated models are exactly the same as results in Ravikumar et al. (2011) with only some minor differences in coefficients. Compare* (31) *with* (17)

for sample complexity and (34) *with* (18) *for convergence rate of estimated Markov matrix* $\widehat{J}_M$*. But regarding the mutual incoherence conditions, we observe that the conditions for the special case sparse inverse covariance estimation in* (16) *are less restrictive than the conditions for the general case in* (12)-(13)*. Since the sparse inverse covariance estimation (Ravikumar et al., 2011) is a special case of the general model in this paper, this additional limitation on models is inevitable, i.e., it is natural that we need some more incoherence conditions in order to be able to recover both the Markov and residual models in the general case.*

5.2 Sparse Covariance Estimation

High-dimensional estimation of sparse covariance models has been studied in Bickel and Levina (2008). They propose an estimation of a class of sparse covariance matrices by "hard thresholding". They also prove spectral norm guarantees on the error between the estimated and exact covariance matrices. We also recover similar results in the other extreme case of proposed program (7) when $\lambda \approx 0$. The program reduces to the sparse covariance estimator as discussed earlier. In order to see that again, let us investigate the dual program restated as follows

$$\begin{aligned}\big(\widehat{\Sigma}_M, \widehat{\Sigma}_R\big) := \underset{\Sigma_M, \Sigma_R}{\arg\max}\ & \log\det \Sigma_M - \lambda \|\Sigma_R\|_{1,\text{off}} \\ \text{s.t. } & \|\widehat{\Sigma}^n - \Sigma_M + \Sigma_R\|_{\infty,\text{off}} \leq \gamma, \\ & \big(\Sigma_M\big)_d = \big(\widehat{\Sigma}^n\big)_d,\ \big(\Sigma_R\big)_d = 0, \\ & \Sigma_M \succ 0, \Sigma_M - \Sigma_R \succ 0.\end{aligned}$$

When the parameter $\lambda \approx 0$, the variable Σ_R is very slightly penalized in the objective function. Therefore, most of the statistical dependencies are captured by Σ_R and thus, off-diagonal entries of Σ_M take very small values. Furthermore, according to the property of optimization program that the support of Σ_R is contained within the support of J_M, sparsity on Σ_R is encouraged by the effect of parameter γ.
It is also observed that we are approximately performing "soft thresholding" in program (7) (when $\lambda \approx 0$) comparing to "hard thresholding" in Bickel and Levina (2008). Consider the case $\lambda = 0$, where the Markov part Σ_M is a diagonal matrix. Therefore, the $\|\widehat{\Sigma}^n - \Sigma_M + \Sigma_R\|_{\infty,\text{off}} \leq \gamma$ constraint in the dual program (7) reduces to $\|\widehat{\Sigma}^n + \Sigma_R\|_{\infty,\text{off}} \leq \gamma$ where it is seen that the negative soft thresholding is performed on matrix $\widehat{\Sigma}^n$ with threshold parameter γ, given by

$$S_\gamma(x) = \text{sign}(-x)(|x| - \gamma)_+.$$

Notice that we need to have $\lambda \approx 0$ for recovering the sparse covariance matrix given empirical covariances and in this case, we can view the estimator as approximately performing soft thresholding.

Finally, we propose the corollary for this special case. Before that, we need some additional definitions for a general covariance matrix Σ^*. Similar to definition (10), the support of a covariance matrix Σ^* is defined as

$$S_\Sigma := \{(i,j) \in V \times V | \Sigma^*_{ij} \neq 0\}.$$

The maximum node degree for a covariance matrix Σ^* is also defined as

$$d_\Sigma := \max_{j=1,\ldots,p} |\{i : (i,j) \in S_\Sigma\}|.$$

Corollary 9 (Sparse Covariance Estimation) *Consider a Gaussian distribution with covariance matrix Σ^* satisfying eigenvalue control condition (A.6). Given a sample covariance matrix $\widehat{\Sigma}^n$ using n i.i.d. samples from the Gaussian model, let $\big(\widehat{\Sigma}_M, \widehat{\Sigma}_R\big)$ denote the optimal solutions of the primal-dual pair (19) and (7), with parameters $\gamma = C_1\sqrt{\log p/n}$ and $\lambda = C_2\sqrt{\log p/n}$ for some constants $C_1, C_2 > 0$. The estimated covariance matrix $\widehat{\Sigma}$ is defined as $\widehat{\Sigma}_{\text{off}} := -\widehat{\Sigma}_R$ and $\widehat{\Sigma}_d := \big(\widehat{\Sigma}_M\big)_d$. Suppose that $\big(\Sigma^*_{\text{off}}\big)_{\min} := \min_{(i,j)\in S_\Sigma, i\neq j} |\big(\Sigma^*\big)_{ij}|$ scales as $\big(\Sigma^*_{\text{off}}\big)_{\min} = \Omega\big(\sqrt{\log p/n}\big)$ and the sample size n is lower bounded as*

$$n = \Omega\big(d_\Sigma^2 \log p\big),$$

then with probability greater than $1 - 1/p^c \to 1$ (for some $c > 0$), we have:

a) The estimate $\widehat{\Sigma}$ satisfies ℓ_∞ bound

$$\|\widehat{\Sigma} - \Sigma^*\|_{\infty,\text{off}} = O\bigg(\sqrt{\frac{\log p}{n}}\bigg).$$

b) The estimate $\widehat{\Sigma}_{\text{off}}$ is sparsistent and sign consistent with Σ^*_{off}.

Proof See Appendix F. ■

5.3 Structured Noise Model

In the discussion up to now, we considered general residual matrices Σ^*_R, not necessarily positive definite, thereby allowing for a rich class of covariance decomposition models. In this section, we modify the conditions and the learning method to incorporate positive-definite residual matrices Σ^*_R.

We regularize the diagonal entries in an appropriate way to ensure that both J^*_M and Σ^*_R are positive definite. Thus, the identifiability assumptions (A.0)-(A.3) are modified as follows:

(A.0') Σ^*, Σ^*_R and J^*_M are positive definite matrices, i.e., $\Sigma^* \succ 0, \Sigma^*_R \succ 0, J^*_M \succ 0$.

(A.1') J^*_M is normalized such that $\big(J^*_M\big)_d = \lambda^*_1$ for some $\lambda^*_1 > 0$ and off-diagonal entries of J^*_M are bounded from above, i.e., $\|J^*_M\|_{\infty,\text{off}} \leq \lambda^*_2$, for some $\lambda^*_2 > 0$.

(A.2') The off-diagonal entries of Σ^*_R satisfy

$$\big(\Sigma^*_R\big)_{ij} \neq 0 \iff |\big(J^*_M\big)_{ij}| = \lambda^*_2, \quad \forall\, i \neq j.$$

(A.3') For any i, j, we have $\text{sign}\big(\big(\Sigma^*_R\big)_{ij}\big).\text{sign}\big(\big(J^*_M\big)_{ij}\big) \geq 0$, i.e, the signs are the same.

It is seen in (A.1') that we put additional restrictions on diagonal entries of the Markov matrix J_M^* in order to have nonzero diagonal entries for the residual matrix Σ_R^*.
Similar to the general form of dual program introduced in (23), we propose the following optimization program to estimate the Markov and residual components in the structured noise model:

$$\begin{aligned}\big(\widehat{\Sigma}_M, \widehat{\Sigma}_R\big) := \underset{\Sigma_M, \Sigma_R \succ 0}{\arg\max} \log\det \Sigma_M - \lambda_1 \|\Sigma_R\|_{1,\text{on}} - \lambda_2 \|\Sigma_R\|_{1,\text{off}} \\ \text{s.t.} \quad \|\widehat{\Sigma}^n + \Sigma_R - \Sigma_M\|_{\infty,\text{off}} \leq \gamma, \\ \big(\widehat{\Sigma}^n\big)_d + \big(\Sigma_R\big)_d = \big(\Sigma_M\big)_d.\end{aligned}$$

The decomposition result under exact statistics can be similarly proven by setting parameter $\gamma = 0$ when the identifiability assumptions (A.0')-(A.3') are satisfied. Furthermore, under additional estimation assumptions (A.4)-(A.6), the sample statistics guarantees in Theorem 5 can be also extended to the solutions of above program.

6. Proof Outline

In this section, the Lagrangian primal form for the proposed dual program (7) is provided first and then the proof outlne is presented. For now, we drop the positive-definiteness constraint $\Sigma_M - \Sigma_R \succ 0$ in the proposed dual program (7). We finally show that this constraint is satisfied for the proposed estimation under specified conditions and thus this constraint can be dropped. In the subsequent discussion, we drop this constraint. It is shown in Appendix A that the primal form for this reduced dual program is

$$\begin{aligned}\widehat{J}_M := \underset{J_M \succ 0}{\arg\min} \ \langle \widehat{\Sigma}^n, J_M \rangle - \log\det J_M + \gamma \|J_M\|_{1,\text{off}} \\ \text{s.t.} \ \|J_M\|_{\infty,\text{off}} \leq \lambda,\end{aligned} \tag{19}$$

We further establish that $\widehat{\Sigma}_M = \widehat{J}_M^{-1}$ is valid between the dual variable Σ_M and primal variable J_M and thus,

$$\|\widehat{\Sigma}^n - \widehat{J}_M^{-1} + \widehat{\Sigma}_R\|_{\infty,\text{off}} \leq \gamma. \tag{20}$$

Comparing the above with the exact decomposition $\Sigma^* = {J_M^*}^{-1} - \Sigma_R^*$ in (2), we note that for the sample version, we do not exactly fit the Markov and the residual models with the sample covariance matrix $\widehat{\Sigma}^n$, but allow for some divergence, depending on γ. Similarly, the primal program (19) has an additional ℓ_1 penalty term on $\widehat{J}_M$, which is absent in (6). Having a non-zero γ in the primal program enables us to impose a sparsity constraint on $\widehat{J}_M$, which in turn, enables us to estimate the matrices in the high dimensional regime ($p \gg n$), under a set of conditions of sufficient conditions given in section 4.2.

We now provide a high-level description of the proof for Theorem 5. The detailed proof is given in Appendix D. The proof is based on the primal-dual witness method, which has been previously employed in Ravikumar et al. (2011) and other works. However, we require significant modifications of this approach in order to handle the more complex setting of covariance decomposition.

In the primal-dual witness method, we define a modified version of the original optimization program (19). Note that the key idea in constructing the modified version is to be

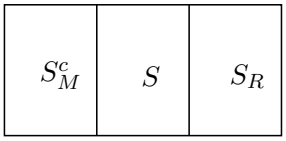


Figure 3: The sets S_R, S and S_M^c form a partition of $\{(1,...,p) \times (1,...,p)\}$, where p is the number of nodes, S_R is the support of the residual covariance matrix Σ_R^* and S_M is the support of the precision matrix J_M^* of the Markov model and S_M^c is its complement.

able to analyze it and prove guarantees for it in a less complicated way comparing to the original version. Let us denote the solutions of the modified program by $\big(\widetilde{J}_M, \widetilde{\Sigma}_R\big)$ pair. In general, the optimal solutions of the two programs, original and modified one, are different. However, under conditions (A.0)–(A.5), we establish that their optimal solutions coincide. See Appendix D for details. Through this equivalence, we thus establish that the optimal solution $\big(\widehat{J}_M, \widehat{\Sigma}_R\big)$ of the original program in (19) inherits all the properties of the optimal solution $\big(\widetilde{J}_M, \widetilde{\Sigma}_R\big)$ of the modified program, i.e., the solutions of the modified program act as witness for the original program. In the following, we define the modified optimization program and its properties. The primal-dual witness method steps which guarantee the equivalence between solutions of the original and the modified program are mentioned in Appendix D.

We modify the sample version of our optimization program in (19) as follows:

$$\widetilde{J}_M := \underset{J_M \succ 0}{\arg\min} \, \langle \widehat{\Sigma}^n, J_M \rangle - \log\det J_M + \gamma \|J_M\|_{1,\text{off}} \qquad (21)$$
$$\text{s.t. } \big(J_M\big)_{S_M^c} = 0, \ \big(J_M\big)_{S_R} = \lambda \operatorname{sign}\Big(\big(J_M^*\big)_{S_R}\Big).$$

Note that since we do not a priori know the supports of the original matrices J_M^* and Σ_R^*, the above program cannot be implemented in practice, but is only a device useful for proving consistency results. We observe that the objective function in the modified program above is the same as the original program in (19), and only the constraints on the precision matrix are different in the two programs. In the above program in (21), constraints on the entries of the precision matrix when limited to sets S_R and S_M^c are more restrictive, while those in set $S := S_M \setminus S_R$ are more relaxed (i.e., the ℓ_∞ constraints present in (19) are removed above), compared to the original program in (19). Recall that S_M denotes the support of the Markov model, while $S_R \subseteq S_M$ denotes the support of the residual or the independence model. See Figure 3.

We now discuss the properties of the optimal solution $\big(\widetilde{J}_M, \widetilde{\Sigma}_R\big)$ of the modified program in (21). Since the precision matrix entries on S_M^c are set to zero in (21), we have that $\text{Supp}(\widetilde{J}_M) \subseteq \text{Supp}(J_M^*)$. Denoting $\widetilde{\Sigma}_R$ as the residual covariance matrix corresponding to the modified program (21), we can similarly characterize it in the following form derived from duality:

$$\big(\widetilde{\Sigma}_R\big)_{ij} = \begin{cases} 0 & \text{for} \quad (i,j) \in S \\ \widetilde{\beta}_{ij} & \text{for} \quad (i,j) \in S_R, S_M^c, \end{cases} \qquad (22)$$

where $\widetilde{\beta}_{ij}$ are the Lagrangian multipliers corresponding to the equality constraints in the modified program (21).

Define estimation errors $\widetilde{\Delta}_J := \widetilde{J}_M - J_M^*$ and $\widetilde{\Delta}_R := \widetilde{\Sigma}_R - \Sigma_R^*$ for the modified program in (21). It is easy to see that $\big(\widetilde{\Delta}_J\big)_{S_R} = \lambda_\delta$, $\big(\widetilde{\Delta}_J\big)_{S_M^c} = 0$, $\big(\widetilde{\Delta}_R\big)_S = 0$, where $\lambda_\delta := \lambda - \lambda^* > 0$. This implies that in any of the three sets S, S_R or S_M^c, only one of the two estimation errors $\widetilde{\Delta}_J$ or $\widetilde{\Delta}_R$ can be non-zero (or is at most λ_δ). This property is crucial to be able to decouple the perturbations in the Markov and the independence domains, and thereby gives bounds on the individual perturbations. It is not clear if there is an alternative partitioning of the variables (here the partition is S, S_R and S_M^c) which allows us to decouple the estimation errors for $\widetilde{J}_M$ and $\widetilde{\Sigma}_R$. Through this decoupling, we are able to provide bounds on estimation errors $\widetilde{\Delta}_J$ and $\widetilde{\Delta}_R$ and thus, Theorem 5 is established.

7. Experiments

In this section, we provide synthetic and real experimental results for the proposed algorithm. We term our proposed optimization program as $\ell_1 + \ell_\infty$ method and compare it with the well-known ℓ_1 method which is a special case of the proposed algorithm when $\lambda = \infty$. The primal optimization program (19) is implemented via the ADMM (Alternating Direction Method of Multipliers) technique proposed in Mohan (2013). We also compare the performance of belief propagation on the proposed model.

7.1 Synthetic Data

We build a Markov + residual synthetic model in the following way. We choose 0.2 fraction of Markov edges randomly to introduce residual edges. The underlying graph for the Markov part is a $q \times q$ 2-D grid structure (4-nearest neighbor grid). Therefore, the number of nodes is $p = q^2$. Because of assumption (A.2), we randomly set 0.2 fraction of nonzero Markov off-diagonal entries to $\{-0.2, 0.2\}$, and the rest of nonzero off-diagonal entries in J_M^* (corresponding to the grid edges) are randomly chosen from set $\pm[0.15, 0.2]$, i.e., $\big(J_M^*\big)_{ij} \in [-0.2, -0.15] \cup [0.15, 0.2]$, for all $(i,j) \in E_M$. Note that 0.2 fraction of edges take the maximum absolute value which is needed by assumption (A.2). Then we ensure that J_M^* is positive definite by adding some uniform diagonal weighting. The nonzero entries of Σ_R^* are chosen from $\pm[0.15, 0.2]$ such that the sign of residual entry is the same as the sign of overlapping Markov entry (assumption (A.3)). We also generate a random mean in the interval $[0, 1]$ for each variable. Note that this generated synthetic model does not necessarily satisfy mutual incoherence conditions (A.4) and (A.5); But we observe in the following that our method has good numerical estimation performance even when the incoherence conditions are not fully satisfied.

Before we provide experiment results, it is worth mentioning that the realization of above model is an example that both Markov and residual matrices J_M^* and Σ_R^* are sparse, while the overall covariance matrix $\Sigma^* = {J_M^*}^{-1} - \Sigma_R^*$ and concentration matrix $J^* = {\Sigma^*}^{-1}$ are both dense matrices.

Size(p)	c_γ	λ
25	2.23	0.2
64	2.08	0.2
100	2.01	0.2
400	1.85	0.2
900	1.83	0.2

Table 1: Regularization parameters used for grid-structured Markov graph simulations in Figure 4. Note that $\gamma = c_\gamma \sqrt{\log p / n}$.

7.1.1 Effect of graph size p

We apply our method ($\ell_1 + \ell_\infty$ method) to random realizations of the above described model $\Sigma^* = {J_M^*}^{-1} - \Sigma_R^*$ with different sizes $p \in \{25, 64, 100, 400, 900\}$. Normalized Dist $\left(\widehat{J}_M, J_M^*\right)$, the edit distance between the estimated and exact Markov components $\widehat{J}_M$ and J_M^*, and normalized Dist $\left(\widehat{\Sigma}_R, \Sigma_R^*\right)$, the edit distance between the estimated and exact residual components $\widehat{\Sigma}_R$ and Σ_R^* as a function of number of samples are plotted in Figure 4 for different sizes p.

In Figure 4.a, normalized Dist $\left(\widehat{J}_M, J_M^*\right)$ is plotted and in Figure 4.b, the same is plotted with rescaled horizontal axis $n/\log p$. We observe that by increasing the number of samples, the edit distance decreases, and by increasing the size of problem, it becomes harder to recover the components which are intuitive. More importantly, we observe in the rescaled graph that the plots for different sizes p make a lineup which is consistent with the theoretical results saying that[8] $n = O(d^2 \log p)$ is sufficient for correct recovery.

Similarly, in Figure 4.c, normalized[9] Dist $\left(\widehat{\Sigma}_R, \Sigma_R^*\right)$ is plotted and in Figure 4.d, the same is plotted with rescaled horizontal axis $n/\log p$. We similarly have the initial observations that by increasing the number of samples, the edit distance decreases, and by increasing the size of problem, it becomes harder to recover the components. The theoretical sample complexity $n = O(d^2 \log p)$ is also validated in Figure 4.d.

The value of regularization parameters used for this simulation are provided in Table 1. Since in the synthetic experiments, we know the value of $\lambda^* := \|J_M^*\|_{\infty,\text{off}}$, parameter λ is set to $\lambda^* = 0.2$. It is observed that the recovery of sparsity pattern of the Markov component J_M^* is fairly robust to the choice of this parameter. For choosing parameter γ, the experiment is run for several values of γ to see which one gives the best recovery result. The effect of parameter γ is discussed in detail in the next subsection.

7.1.2 Effect of regularization parameter γ

We apply our method ($\ell_1 + \ell_\infty$ method) to random realizations of the above described grid-structured synthetic model $\Sigma^* = {J_M^*}^{-1} - \Sigma_R^*$ with fixed size $p = 64$. Here, we fix the

8. Note that in the grid graph, $d = 4$ is fixed for different sizes p.
9. The normalized distance for recovering residual component is greater than 1 for small n. Since we normalize the distance with the number of edges in the exact model, this may happen.

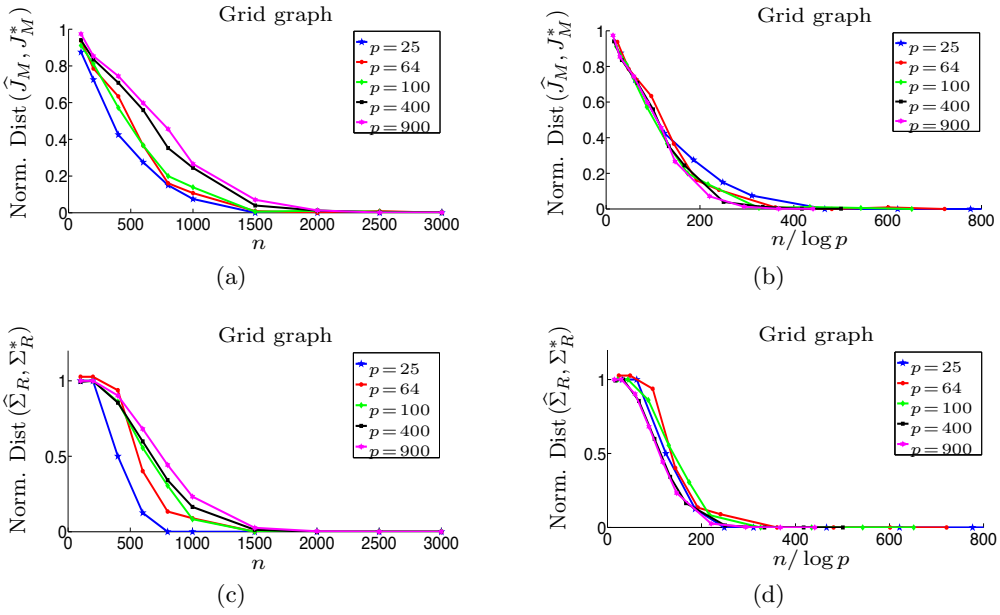


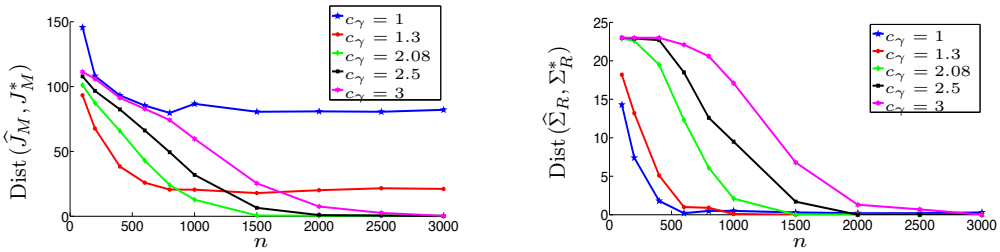
Figure 4: Simulation results for grid-structured Markov graph with different size p. (a-b) Normalized edit distance between the estimated Markov component $\widehat{J}_M$ and the exact Markov component J^*_M. In panel (b), the horizontal axis is rescaled as $n/\log p$. (c-d) Normalized edit distance between the estimated residual component $\widehat{\Sigma}_R$ and the exact residual component Σ^*_R. In panel (d), the horizontal axis is rescaled as $n/\log p$. Each point in the figures is derived from averaging 10 trials.

regularization parameter[10] $\lambda = 0.2$ and change the regularization parameter $\gamma = c_\gamma \sqrt{\log p/n}$ where $c_\gamma \in \{1, 1.3, 2.08, 2.5, 3\}$. The edit distance between the estimated and exact Markov components $\widehat{J}_M$ and J^*_M, and the edit distance between the estimated and exact residual components $\widehat{\Sigma}_R$ and Σ^*_R are plotted in Figure 5. We observe the pattern that for c_γ less than some optimal value c^*_γ, the Markov component is not recovered, and for values greater than the optimal value, the components are recovered with different statistical efficiency, where by increasing c_γ, the statistical rate of Markov component recovery becomes worse. For the simulations of previous subsection provided in Figure 4, we choose some regularization parameter close to c^*_γ. For example, we choose $c_\gamma = 2.08$ for $p = 64$ as suggested by Figure 5.

7.1.3 Comparing $\ell_1 + \ell_\infty$ and ℓ_1 methods

We apply $\ell_1 + \ell_\infty$ and ℓ_1 methods to a random realization of the above described grid-structured synthetic model[11] $\Sigma^* = {J^*_M}^{-1} - \Sigma^*_R$ with size $p = 64$. The edit distance between the estimated and exact Markov components $\widehat{J}_M$ and J^*_M is plotted in Figure 6.a. We observe that the behaviour of $\ell_1 + \ell_\infty$ method is very close to ℓ_1 method which suggests that

10. λ is set to the maximum absolute value of off-diagonal entries of Markov matrix J^*_M.
11. Here, we choose the nonzero off-diagonal entries of J^*_M randomly from $\{-0.2, 0.2\}$.



(a) (b)

Figure 5: Simulation results for grid graph with fixed size $p = 64$ and regularization parameters $\lambda = 0.2$ and varying $c_\gamma \in \{1, 1.3, 2.08, 2.5, 3\}$ where $\gamma = c_\gamma\sqrt{\log p/n}$. (a) Edit distance between the estimated Markov component $\widehat{J}_M$ and the exact Markov component J^*_M. (b) Edit distance between the estimated residual component $\widehat{\Sigma}_R$ and the exact residual component Σ^*_R. Each point in the figures is derived from averaging 10 trials.

sparsity pattern of J^*_M can be estimated efficiently under either methods. The edit distance between the estimated and exact residual components $\widehat{\Sigma}_R$ and Σ^*_R is plotted in Figure 6.b. Since there is not any off-diagonal ℓ_∞ constraints in ℓ_1 method, it can not recover the residual matrix Σ^*_R. Finally the ℓ_∞-elementwise norm of error between the estimated precision matrix $\widehat{J}$ and the exact precision matrix J^* is sketched for both methods in Figure 6.c. We observe the advantage of proposed $\ell_1 + \ell_\infty$ method in estimating the overall model precision matrix $J^* = \Sigma^{*-1}$. Note that the same regularization parameters provided in Table 1 are used for the simulations of this subsection, except for ℓ_1 method that we have $\lambda = \infty$.

7.1.4 Benefit of applying LBP (Loopy Belief Propagation) to the proposed model

We compare the result of applying LBP to J^* and J^*_M components of a random realization of the above described grid-structured synthetic model.[12] The log of average mean and variance errors over all nodes are sketched in Figure 7 throughout the iterations. We observe that LBP does not converge for J^* model. It is shown in Malioutov et al. (2006) that if a model is walk-summable, then the mean estimates under LBP converge and are correct. The spectral norms of the partial correlation matrices are $\|\overline{R}_M\| = 0.8613$ and $\|\overline{R}\| = 3.2446$ for J^*_M and J^* models respectively. Thus, the matrix J^* is not walk-summable and therefore its convergence under LBP is not guaranteed and this is seen in Figure 7. On the other hand, LBP is accurate for J^*_M matrix. Thus, our method learns models which are better suited for inference under loopy belief propagation.

7.2 Real Data

The proposed algorithm is also applied to foreign exchange rate and monthly stock returns data sets to learn a Markov plus residual model introduced in the paper. It is important to

12. Here, we choose 0.5 fraction of Markov edges randomly to introduce residual edges.

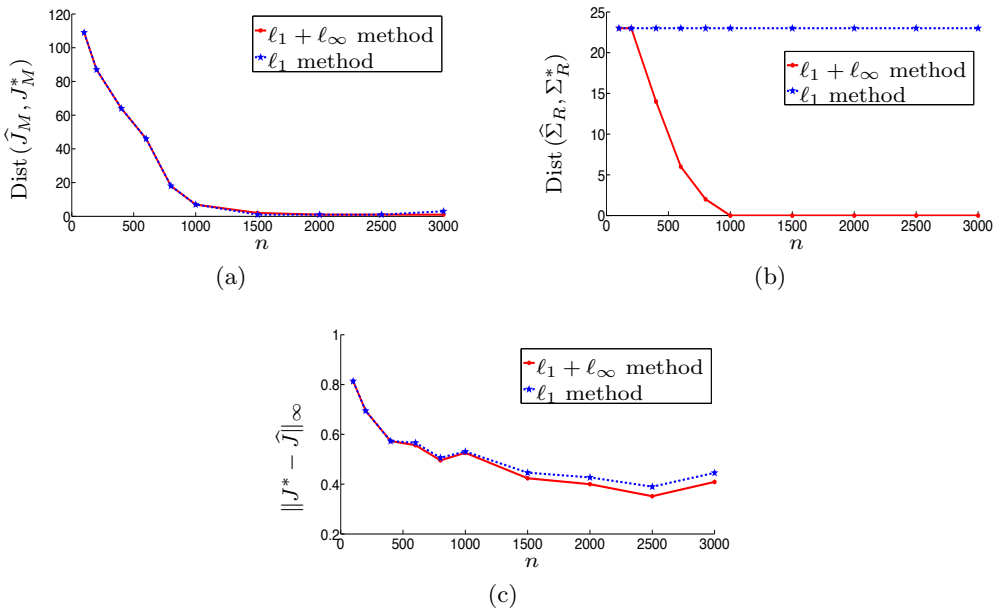


Figure 6: Simulation results for grid graph with size $p = 64$. (a) Edit distance between the estimated Markov component $\widehat{J}_M$ and the exact Markov component J^*_M. (b) Edit distance between the estimated residual component $\widehat{\Sigma}_R$ and the exact residual component Σ^*_R. (c) Precision matrix estimation error $\|J^* - \widehat{J}\|_\infty$, where $\widehat{J} = \widehat{J}_M$ for ℓ_1 method and $\widehat{J} = \left(\widehat{J}_M^{-1} - \widehat{\Sigma}_R\right)^{-1}$ for $\ell_1 + \ell_\infty$ method.

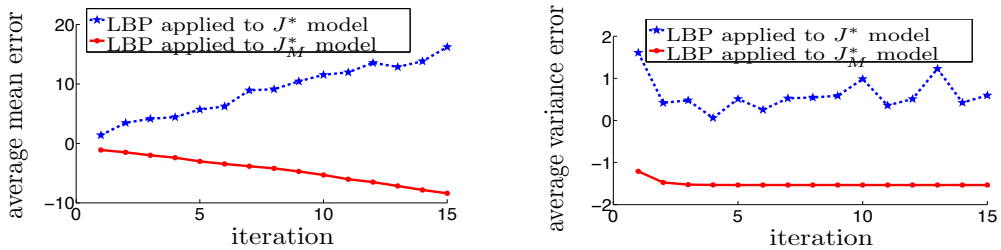


Figure 7: Performance under loopy belief propagation for the overall model (J^*) and the Markov component (J^*_M).

note that the real data sets can be modeled by different models not necessarily satisfying the conditions proposed in this paper. But, here we observe that the resulting Markov plus residual models are fairly interpretable for the corresponding real data sets. The interpretations are discussed in detail in the following sections.

7.2.1 Foreign exchange rate data

In this section, we apply the proposed algorithm to the foreign exchange rate data set.[13] The data set includes monthly exchange rates of 19 countries currency with respect to US dollars from October 1983 to January 2012. Thus, the data set has 340 samples of 19 variables. We apply the optimization program (7) with a slight modification. Since the underlying model for this data set does not necessarily satisfy the proposed eigenvalue condition (A.6), we need to make sure that the overall covariance matrix estimation $\widehat{\Sigma}$ is positive definite and thus a valid covariance matrix. We add an additional constraint to the optimization program (7), imposing a lower bound on the minimum eigenvalue of overall covariance matrix $\lambda_{min}(\Sigma)$, i.e., $\lambda_{min}(\Sigma) \geq \sigma_{min}$. The parameter σ_{min} is set to 0.001 in this experiment.

The resulting edges of Markov and residual matrices for some moderate choice of regularization parameters $\gamma = 20$ and $\lambda = 0.004$ are plotted in Figure 8. The choice of regularization parameters are further discussed at the end of this subsection. We observe sparsity on both Markov and residual structures. There are two main observations in the learned model in Figure 8. First, it is seen that the statistical dependencies of foreign exchange rates are correlated with the geographical locations of countries, e.g., it is observed in the learned model that the exchange rates of Asian countries are more correlated. We can refer to Asian countries "South Korea", "Japan","China","Sri Lanka", "Taiwan", "Thailand" and "India" in the Markov model where several edges exist between them while other nodes in the graph have much lower degrees. We observe similar patterns in the residual matrix, e.g., there is an edge between "India" and "Sri Lanka" in the residual model. We also see the interesting phenomena in the Markov graph that there exist some high degree nodes such as "South Korea" and "Japan". The presence of high degree nodes suggests that incorporating hidden variables can further lead to sparser representations, and this has been observed before in other works, e.g., Choi et al. (2010); Chandrasekaran et al. (2010a); Choi et al. (2011).

The regularization parameters are chosen such that the resulting Markov and residual graphs are reasonably sparse, while still being informative. Increasing the parameter γ makes both Markov and residual components sparser, and increasing parameter λ makes the residual component sparser. In addition, it is worth discussing the fact that we chose parameter γ relatively large compared to parameter λ in this simulation. In Theorem 4, we have $\gamma = C_1\sqrt{\log p/n}$ and $\lambda = \lambda^* + C_2\sqrt{\log p/n}$. Now, if C_1 is large compared to C_2 and furthermore λ^* is small, γ can be larger than λ. Hence, we have an agreement between theory and practice.

7.2.2 Monthly stock returns data

In this section we apply the algorithm to monthly stock returns of a number of companies in the S&P 100 stock index. We pick 17 companies in divisions "E.Trans, Comm, Elec&Gas" and "G.Retail Trade" and apply the optimization program (19) to their stock returns data to learn the model. The resulting edges for Markov and residual matrices are plotted in Figure 9 for regularization parameters $\gamma = 2.2e-03$ and $\lambda = 1e-04$. There is sparsity on both Markov and residual structure. The isolated nodes in the Markov graph are not

13. Data set available at `http://research.stlouisfed.org/fred2/categories/15/downloaddata`.

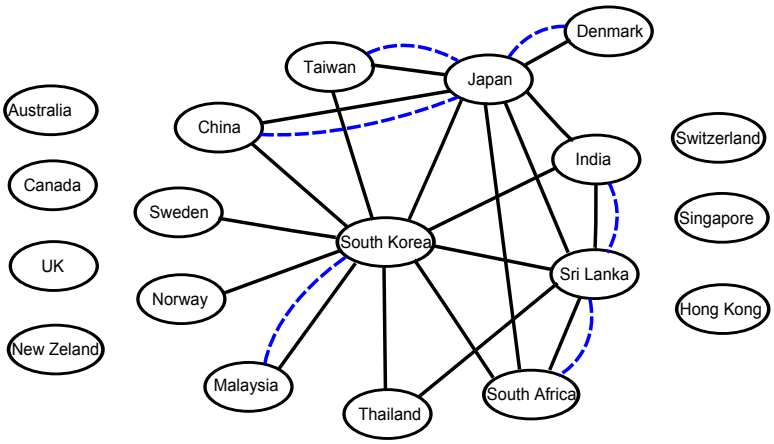


Figure 8: Markov and independence graph structures for the foreign exchange rate data set with regularization parameters $\gamma = 20$ and $\lambda = 0.004$. Solid edges indicate Markov model and dotted edges indicate independence model.

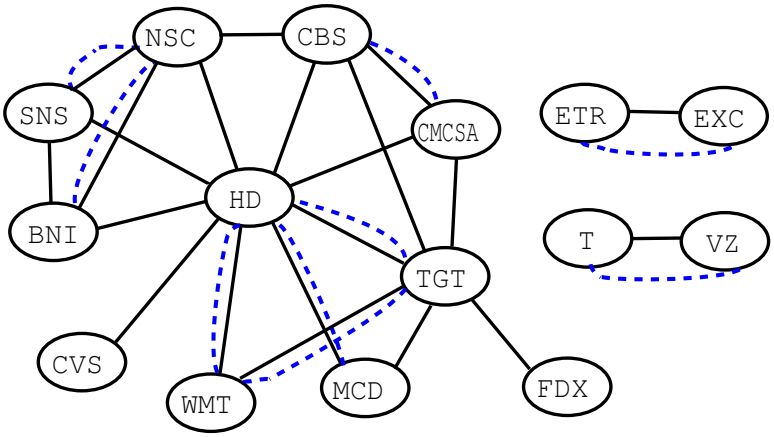


Figure 9: Markov and independence graph structures for the monthly stock returns data set with regularization parameters $\gamma = 2.2e-03$ and $\lambda = 1e-04$. Solid edges indicate Markov model and dotted edges indicate independence model.

presented in the figure. We see in both Markov and residual graphs that there exist higher correlations among stock returns of companies in the same division or industry. There are 5 connected partitions in the residual graph. e.g., nodes "HD", "WMT", "TGT" and "MCD", all belonging to division Retail Trade form a partition. This is also observed for the telecommunication industries (companies "T" and "VZ") and energy industries (companies "ETR" and "EXC"). We see a similar pattern in the Markov graph but with more edges. Similar to exchange rate data set results, we also observe high degree nodes in the Markov graph such as "HD" and "TGT" which suggest incorporating hidden nodes.

8. Conclusion

In this paper, we provided an in-depth study of convex optimization methods and guarantees for high-dimensional covariance matrix decomposition. Our methods unify the existing results for sparse covariance/precision estimation and introduce a richer class of models with sparsity in multiple domains. We provide consistency guarantees for estimation in both the Markov and the residual domains, and establish efficient sample complexity results for our method. These findings open up many future directions to explore. One important aspect is to relax the sparsity constraints imposed in the two domains, and to develop new methods to enable decomposition of such models. Other considerations include extension to discrete models and other models for the residual covariance matrix (e.g., low rank matrices). Such findings will push the envelope of efficient models for high-dimensional estimation. It is worth mentioning while in many scenarios it is important to incorporate latent variables, in our framework it is challenging to incorporate both latent variables as well as marginal independencies, and provide learning guarantees, and we defer it to future work.

Acknowledgements

We thank Karthik Mohan for helpful discussions on running experiments. We also acknowledge useful discussions with Max Welling, Babak Hassibi and Martin Wainwright. We also thank Bin Yu and the JMLR reviewers for valuable comments that have significantly improved the manuscript. M. Janzamin is supported by NSF Award CCF-1219234 and ARO Award W911NF-12-1-0404. A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, NSF Award CCF-1219234, AFOSR Award FA9550-10-1-0310, and ARO Award W911NF-12-1-0404.

Appendix A. Duality Between Programs

In this section we prove duality between programs (19) and (7) (when the positive-definiteness constraint $\Sigma_M - \Sigma_R \succ 0$ is dropped). By doing this, the duality between programs (6) and (4) is also proved since they are special cases of (19) and (7) when γ is set to zero and $\widehat{\Sigma}^n$ is substituted with Σ^*.

Before we prove duality, we introduce the concept of *subdifferential* or *subgradient* for a convex function not necessarily differentiable. Subgradient (subdifferential) generalizes the gradient (derivative) concept to nondifferentiable functions. Supposing convex function $f : \mathbb{R}^n \to \mathbb{R}$, the subgradient at a point x_0 which is usually denoted by $\partial f(x_0)$ consists of all vectors c such that

$$f(x) \geq f(x_0) + \langle c, x - x_0 \rangle, \quad \forall x \in \mathrm{Dom}\, f.$$

In order to prove duality, we start from program (7) (when the positive-definiteness constraint $\Sigma_M - \Sigma_R \succ 0$ is dropped) and derive the primal form (19). Program (7) can be written in the following equivalent form where λ_1 goes to infinity and λ_2 is used instead of λ.

$$\big(\widehat{\Sigma}_M, \widehat{\Sigma}_R\big) := \underset{\Sigma_M \succ 0, \Sigma_R}{\arg\max} \log\det \Sigma_M - \lambda_1 \|\Sigma_R\|_{1,\mathrm{on}} - \lambda_2 \|\Sigma_R\|_{1,\mathrm{off}} \tag{23}$$

$$\text{s.t.} \quad \|\widehat{\Sigma}^n - \Sigma_M + \Sigma_R\|_{\infty,\text{off}} \leq \gamma,$$
$$\big(\Sigma_M\big)_d - \big(\Sigma_R\big)_d = \big(\widehat{\Sigma}^n\big)_d.$$

By introducing the dual variable J_M for above program, we have:

$$\min_{\substack{\|J_M\|_{\infty,\text{on}}\leq\lambda_1 \\ \|J_M\|_{\infty,\text{off}}\leq\lambda_2}} -\langle J_M, \Sigma_R\rangle = -\lambda_1\|\Sigma_R\|_{1,\text{on}} - \lambda_2\|\Sigma_R\|_{1,\text{off}},$$

where $(\widehat{J}_M)_{\text{on}} \in \lambda_1\partial\|\widehat{\Sigma}_R\|_{1,\text{on}}$, $(\widehat{J}_M)_{\text{off}} \in \lambda_2\partial\|\widehat{\Sigma}_R\|_{1,\text{off}}$ minimizes the above program. Thus, we have the following equivalent form for program (23):

$$\min_{\substack{\|J_M\|_{\infty,\text{on}}\leq\lambda_1 \\ \|J_M\|_{\infty,\text{off}}\leq\lambda_2}} \max_{\substack{\Sigma_M\succ 0,\Sigma_R \\ \|\widehat{\Sigma}^n-\Sigma_M+\Sigma_R\|_{\infty,\text{off}}\leq\gamma \\ (\Sigma_M)_d-(\Sigma_R)_d=\left(\widehat{\Sigma}^n\right)_d}} \log\det\Sigma_M - \langle J_M, \Sigma_R\rangle,$$

where the order of programs is exchanged. If we define the new variable $\Sigma = \Sigma_M - \Sigma_R$, and use Σ as the new variable in the program instead of Σ_R, the inner max program becomes

$$\max_{\substack{\Sigma_M\succ 0,\Sigma \\ \|\widehat{\Sigma}^n-\Sigma\|_{\infty,\text{off}}\leq\gamma,\Sigma_d=\left(\widehat{\Sigma}^n\right)_d}} \log\det\Sigma_M - \langle J_M, \Sigma_M\rangle + \langle J_M, \Sigma\rangle.$$

Since the objective function and constraints are disjoint functions of variables Σ and Σ_M, we can do optimization individually for two variables. The optimizers are $\widehat{\Sigma}_M = J_M^{-1}$ and $\widehat{\Sigma} = \widehat{\Sigma}^n + \gamma Z_\gamma$, where Z_γ is a member of the subgradient of $\|\cdot\|_{1,\text{off}}$ evaluated at point J_M, i.e.,

$$(Z_\gamma)_{ij} = \begin{cases} 0 & \text{for} \quad i = j \\ \in [-1,1] & \text{for} \quad i \neq j, \big(J_M\big)_{ij} = 0 \\ \text{sign}\big(\big(J_M\big)_{ij}\big) & \text{for} \quad i \neq j, \big(J_M\big)_{ij} \neq 0. \end{cases}$$

Also note that since Σ_M should be positive definite, the variable J_M should be also positive definite. Therefore, it adds another constraint $J_M \succ 0$. If we substitute these optimizers, we get the dual program

$$\min_{\substack{J_M\succ 0 \\ \|J_M\|_{\infty,\text{on}}\leq\lambda_1 \\ \|J_M\|_{\infty,\text{off}}\leq\lambda_2}} \langle\widehat{\Sigma}^n, J_M\rangle - \log\det J_M + \gamma\|J_M\|_{1,\text{off}},$$

which is equivalent to (19) when λ_1 goes to infinity and therefore the result is proved.

Appendix B. Characterization of the Proposed Optimization Programs

We proposed programs (6) and (19) to do decomposition and estimation respectively. Former is used to decompose exact statistics to its Markov and residual covariance components and the latter is used to estimate decomposition components given sample covariance matrix. In this appendix we characterize optimal solutions of these optimization programs. Both programs are convex and therefore the optimal solutions can be characterized using standard convex optimization theory. Note that the proof of following lemmas is mentioned after the remarks.

Lemma 1 *For any $\lambda > 0$, primal problem (6) has a unique solution $\widehat{J}_M \succ 0$ which is characterized by the following equation:*

$$\Sigma^* - \widehat{J}_M^{-1} + \widehat{Z} = 0, \tag{24}$$

where $\widehat{Z}$ has the following form

$$\widehat{Z}_{ij} = \begin{cases} 0 & \text{for} \quad i = j \\ 0 & \text{for} \quad i \neq j, |\big(\widehat{J}_M\big)_{ij}| < \lambda \\ \widehat{\alpha}_{ij} \operatorname{sign}\big(\big(\widehat{J}_M\big)_{ij}\big) & \text{for} \quad i \neq j, |\big(\widehat{J}_M\big)_{ij}| = \lambda, \end{cases} \tag{25}$$

in which $\widehat{\alpha}_{ij}$ can only take nonnegative values, i.e., we have $\widehat{\alpha}_{ij} \geq 0$.

Remark 10 *Comparing Lagrangian optimality condition in (24) with relation $\Sigma^* = \widehat{J}_M^{-1} - \widehat{\Sigma}_R$ between solutions of primal-dual optimization programs (derived in Appendix A) implies the equality $\widehat{\Sigma}_R = \widehat{Z}$. Thus, $\widehat{\Sigma}_R$ entries are determined by Lagrangian multipliers of primal program. More specifically, we have*

$$(\widehat{\Sigma}_R)_{ij} = \begin{cases} 0 & \text{for} \quad i = j \\ 0 & \text{for} \quad i \neq j, |\big(\widehat{J}_M\big)_{ij}| < \lambda \\ \widehat{\alpha}_{ij} \operatorname{sign}\big(\big(\widehat{J}_M\big)_{ij}\big) & \text{for} \quad i \neq j, |\big(\widehat{J}_M\big)_{ij}| = \lambda, \end{cases} \tag{26}$$

where $\widehat{\alpha}_{ij} \geq 0$ are the Lagrangian multipliers of primal program (6).

Lemma 2 *For any $\lambda > 0$, $\gamma \geq 0$ and sample covariance matrix $\widehat{\Sigma}^n$ with strictly positive diagonal entries, primal problem (19) has a unique solution $\widehat{J}_M \succ 0$ which is characterized by the equation*

$$\widehat{\Sigma}^n - \widehat{J}_M^{-1} + \widehat{Z} = 0, \tag{27}$$

where $\widehat{Z} = \widehat{Z}_\alpha + \gamma \widehat{Z}_\gamma$. Matrix $\widehat{Z}_\gamma \in \partial \|\widehat{J}_M\|_{1,\text{off}}$ and $\widehat{Z}_\alpha$ is represented as in (25) for some Lagrangian multipliers $\widehat{\alpha}_{ij} \geq 0$.

Remark 11 *Comparing Lagrangian optimality condition in (27) with relation $\widehat{\Sigma}^n = \widehat{J}_M^{-1} - \widehat{\Sigma}_R - \gamma \widehat{Z}_\gamma$ between solutions of primal-dual optimization programs (derived in Appendix A) implies the equality $\widehat{\Sigma}_R = \widehat{Z}_\alpha$. Thus, $\widehat{\Sigma}_R$ entries are determined by the Lagrangian multipliers of primal program. More specifically, we have*

$$(\widehat{\Sigma}_R)_{ij} = \begin{cases} 0 & \text{for} \quad i = j \\ 0 & \text{for} \quad i \neq j, |\big(\widehat{J}_M\big)_{ij}| < \lambda \\ \widehat{\alpha}_{ij} \operatorname{sign}\big(\big(\widehat{J}_M\big)_{ij}\big) & \text{for} \quad i \neq j, |\big(\widehat{J}_M\big)_{ij}| = \lambda, \end{cases} \tag{28}$$

where $\widehat{\alpha}_{ij} \geq 0$ are the Lagrangian multipliers of primal program (19).

Proof We prove Lemma 2 here and Lemma 1 is a special case of that when γ is set to zero and $\widehat{\Sigma}^n$ is substituted with Σ^*.

For any $\lambda > 0$ and $\gamma \geq 0$, the optimization problem (19) is a convex programming where the objective function is strictly convex. Therefore, if the minimum is achieved it

is unique. Since off-diagonal entries of J_M are bounded according to constraints, the only issue for minimum achievement may arises for unbounded diagonal entries. It is shown in Ravikumar et al. (2011) that if diagonal entries of $\widehat{\Sigma}^n$ are strictly positive, the function is coercive with respect to diagonal entries and therefore here is no issue regarding unbounded diagonal entries. Thus, the minimum is attained in $J_M \succeq 0$. But since when J_M approaches the boundary of positive definite cone, the objective function goes to infinity, the solution is attained in the interior of the cone $J_M \succ 0$. After showing that the unique minimum is achieved, let us characterize the minimum.

Considering α_{ij} as Lagrangian multipliers of inequality constraints of program (19), the Lagrangian function is

$$\mathcal{L}(J_M, \alpha) = \langle \widehat{\Sigma}^n, J_M \rangle - \log\det J_M + \gamma \|J_M\|_{1,\text{off}} + \sum_{i \neq j} \alpha_{ij} \big[\big| \big(J_M\big)_{ij} \big| - \lambda \big].$$

We skipped positive definiteness constraint in writing Lagrangian function since it is inactive. Based on standard convex optimization theory, the matrix $\widehat{J}_M \succ 0$ is the optimal solution if and only if it satisfies KKT conditions. It should minimize the Lagrangian which happens if and only if 0 belongs to the subdifferential of Lagrangian or equivalently there exists a matrix $\widehat{Z}$ such that

$$\widehat{\Sigma}^n - \widehat{J}_M^{-1} + \widehat{Z} = 0,$$

where $\widehat{Z} = \widehat{Z}_\alpha + \gamma \widehat{Z}_\gamma$. Matrix $\widehat{Z}_\gamma \in \partial \|\widehat{J}_M\|_{1,\text{off}}$ and $\widehat{Z}_\alpha$ is

$$(\widehat{Z}_\alpha)_{ij} = \begin{cases} 0 & \text{for} \quad i = j \\ \in \widehat{\alpha}_{ij}.[-1,1] & \text{for} \quad i \neq j, \big(\widehat{J}_M\big)_{ij} = 0 \\ \widehat{\alpha}_{ij} \operatorname{sign}\big(\big(\widehat{J}_M\big)_{ij}\big) & \text{for} \quad i \neq j, \big(\widehat{J}_M\big)_{ij} \neq 0, \end{cases}$$

for some Lagrangian multipliers $\widehat{\alpha}_{ij} \geq 0$. The solution should also satisfy complementary slackness conditions $\widehat{\alpha}_{ij}.\big[\big|\big(\widehat{J}_M\big)_{ij}\big| - \lambda\big] = 0$ for $i \neq j$. Applying this condition to above $\widehat{Z}_\alpha$ representation, results to (25) form proposed in the lemma. ■

Appendix C. Proof of Theorem 4

First note that as mentioned in Remark 3, the pair $\big(\widehat{J}_M, \widehat{\Sigma}_R\big)$ given by optimization program gives a decomposition $\Sigma^* = \widehat{J}_M^{-1} - \widehat{\Sigma}_R$ which is desired.

Next, in order to prove the equivalence, we show that there is a one to one correspondence between the specified conditions (A.0)-(A.3) for valid decomposition and the characterization of optimal solution of optimization program given in lemma 1. We go through each of these conditions one by one in the following lines.

Condition (A.0) is considered in optimization program as positive definiteness of Markov matrix J_M.

Condition (A.1) is exactly the primal constraint $\|J_M^*\|_{\infty,\text{off}} \leq \lambda$.

Condition (A.2) is exactly the same as relation (26) where diagonal entries of residual covariance matrix are zero and its off-diagonal entries can be nonzero only if the absolute

value of corresponding entry in Markov matrix takes the maximum value λ.
Condition (A.3) is exactly the same as inequality $\widehat{\alpha}_{ij} \geq 0$.
In the above lines, we covered one by one correspondence for conditions (A.0)-(A.3). But note that we also covered all the equalities and inequalities that characterize unique optimal solution of optimization program. In other words by above correspondence we proved that both of the following derivations are true where second one is the reverse of first one. On one hand, any optimal solution of optimization program gives a valid decomposition under desired conditions. On the other hand, any valid decomposition under desired conditions is a solution of proposed optimization program. Thus, we can infer that these two are exactly equivalent and the result is proved. Since the solution of optimization program is unique and according to the equivalence between this solution and decomposition under those conditions, uniqueness is also established. ■

Appendix D. Proof of Theorem 5

In this appendix, we first mention an outline of the primal-dual witness method and then provide the detailed proof of the theorem.

D.1 Primal-Dual Witness Method

First, continuing the proof outline presented in section 6, we provide an outline of the primal-dual witness method steps in order to establish equivalence between optimal solutions of the original (19) and the modified (21) optimization programs.

1. The primal witness matrix $\widetilde{J}_M$ is defined as in (21).
2. The dual witness matrix is set as $\widetilde{Z} = -\widehat{\Sigma}^n + \widetilde{J}_M^{-1}$. It is defined in this way to satisfy original program optimal solution characterization mentioned in appendix B.
3. We need to check the following feasibility conditions under which the modified program solution is equivalent to the solution of original one:
   (a) $\|\widetilde{J}_M\|_{\infty,\text{off},S} \leq \lambda$: Since we relaxed the ℓ_∞ bounds on off-diagonal entries in set S, we need to make sure that the modified solution satisfies this bound in order to have equivalence between modified and original programs solutions.
   (b) Set $\big(\widetilde{Z}_\alpha\big)_{S_R} = \big(-\widehat{\Sigma}^n + \widetilde{J}_M^{-1} - \gamma\big(\widetilde{Z}_\gamma\big)\big)_{S_R}$ where $\widetilde{Z}_\gamma \in \partial\|\widetilde{J}_M\|_{1,\text{off}}$. Note that since $|\big(\widetilde{J}_M\big)_{ij}| = \lambda \neq 0$ for any $(i,j) \in S_R$, then $\widetilde{Z}_\gamma$ and therefore above equation is well-defined. Now we need to check: $\big(\widetilde{Z}_\alpha\big)_{ij}\big(\widetilde{J}_M\big)_{ij} \geq 0$ for all $(i,j) \in S_R$. This means that they have the same sign or one of them is zero. We need this condition for equivalence between solutions because Lagrangian multipliers in original program (19) corresponding to inequality constraints should be nonnegative.
   (c) $\|\widetilde{Z}\|_{\infty,S_M^c} < \gamma$: According to the $\big(J_M\big)_{S_M^c} = 0$ constraint in the modified program, all the inequality constraints become inactive in the original one when desired $\widehat{J}_M = \widetilde{J}_M$ equality is satisfied. Then, complementary slackness condition enforce all the Lagrangian multipliers corresponding to set S_M^c to be zero. These can be satisfied by the above strict dual feasibility. Also note that having

zero Lagrangian multipliers results in zero residual entries, i.e., $\big(\widetilde{\Sigma}_R\big)_{S_M^c} = 0$ and therefore $\|\widetilde{\Delta}_R\|_{\infty,S_M^c} = 0$ when this feasibility condition is satisfied.

Also note that we dropped the positive-definiteness constraint $\Sigma_M - \Sigma_R \succ 0$ in the proof outline. Thus, in addition to above conditions, we also need to show that $\widetilde{\Sigma} = \widetilde{\Sigma}_M - \widetilde{\Sigma}_R \succ 0$ in the modified program.

Before we state the detailed proof for the theorem, we introduce a pair of definitions which are used in the analysis. Let us define matrix E as difference between sample covariance matrix and the exact covariance matrix

$$E := \widehat{\Sigma}^n - \Sigma^*. \tag{29}$$

We also define $R\big(\widetilde{\Delta}_J\big)$ as the difference between $\widetilde{J}_M^{-1}$ and its first order Taylor expansion around J_M^*. Recall that $\widetilde{\Delta}_J$ was defined as $\widetilde{\Delta}_J := \widetilde{J}_M - J_M^*$. According to results for first order derivative of inverse function J_M^{-1} (Boyd and Vandenberghe, 2004), the remainder is

$$R\big(\widetilde{\Delta}_J\big) = \widetilde{J}_M^{-1} - {J_M^*}^{-1} + {J_M^*}^{-1}\widetilde{\Delta}_J {J_M^*}^{-1}. \tag{30}$$

D.2 Proof of the Theorem

Exploiting lemmata mentioned in Appendix E, Theorem 5 is proved as follows:

Proof According to the sample error bound mentioned in Lemma 4, we have $\|E\|_\infty \leq \overline{\delta}_f(p^\tau; n)$ for some $\tau > 2$ with probability greater than or equal to $1 - 1/p^{\tau-2}$. In the discussion after this, it is assumed that the above bound for $\|E\|_\infty$ is satisfied and therefore the following results are valid with probability greater than or equal to $1 - 1/p^{\tau-2}$.

By choosing $\gamma = \frac{m}{\alpha}\overline{\delta}_f(p^\tau; n)$, we have $\|E\|_\infty \leq \frac{\alpha}{m}\gamma$ as desired for Lemma 5. Choosing λ_δ as in (36) (compatible with what mentioned in the theorem), we only need to show that the other bound on $\|R\|_\infty$ is also satisfied to be able to apply Lemma 5. As stated in the remark after Theorem 5, the bound on sample complexity is not asymptotic and we assume the following lower bound on the number of samples which is compatible with the asymptotic form mentioned in the theorem:

$$n > \overline{n}_f\left(p^\tau; 1/\max\left\{v_*, 4ld\Big(1+\frac{m}{\alpha}\Big)K_{SS}K_M \max\Big\{1, \frac{4}{l-1}\Big(1+\frac{m}{\alpha}\Big)K_{SS}K_M^2\Big\}\right\}\right), \tag{31}$$

for some $l > 1$. Because of monotonic behaviour of the tail function, for any n satisfying above bound, we have:

$$\overline{\delta}_f\big(p^\tau; n\big) \leq \min\left\{\frac{1}{v_*}, \frac{1}{4ld(1+\frac{m}{\alpha})K_{SS}K_M}, \frac{l-1}{16ld(1+\frac{m}{\alpha})^2K_{SS}^2K_M^3}\right\}, \tag{32}$$

According to the selection for regularization parameters λ_δ and γ and the bound on sample error $\|E\|_\infty$, we have:

$$r := 2K_{SS_R}\lambda_\delta + 2K_{SS}\big(\|E\|_\infty + \gamma\big) \leq \left[\frac{4K_{SS_R}K_{SS}}{1-2K_{SS_R}}\Big(1+\frac{\alpha}{m}\Big)\frac{m}{\alpha} + 2K_{SS}\Big(1+\frac{m}{\alpha}\Big)\right]\overline{\delta}_f(p^\tau; n)$$

$$
\begin{aligned}
&= 2K_{SS}\Big(1+\frac{m}{\alpha}\Big)\overline{\delta}_f(p^\tau;n)\frac{1}{1-2K_{SS_R}} \quad (=:\lambda_\delta)\\
&< 4K_{SS}\Big(1+\frac{m}{\alpha}\Big)\overline{\delta}_f(p^\tau;n),
\end{aligned}
$$

where in the last inequality, we used the second condition is assumption (A.4) that $K_{SS_R} < 1/4$. Note that second line is equal to λ_δ since we assigned the same value in (36). Applying the bound (32) on above inequality, we have

$$
\begin{aligned}
2K_{SS_R}\lambda_\delta + 2K_{SS}\big(\|E\|_\infty+\gamma\big) &< \min\left\{\frac{1}{ldK_M}, \frac{l-1}{4ld(1+\frac{m}{\alpha})K_{SS}K_M^3}\right\}\\
&\le \min\left\{\frac{1}{ldK_M}, \frac{l-1}{2ldK_{SS}K_M^3}\right\}.
\end{aligned}
$$

Thus, the conditions for Lemma 7 are satisfied and we have

$$
\|\widetilde{\Delta}_J\|_{\infty,S} \le 2K_{SS_R}\lambda_\delta + 2K_{SS}\big(\|E\|_\infty+\gamma\big) \le \lambda_\delta < 4K_{SS}\Big(1+\frac{m}{\alpha}\Big)\overline{\delta}_f(p^\tau;n). \tag{33}
$$

Above inequalities tell us multiple things. First, since the error $\|\widetilde{\Delta}_J\|_{\infty,S}$ is bounded by λ_δ, the $\widetilde{J}_M$ entries in set S can not deviate from exact one J_M^* more than λ_δ. We also assumed that the off-diagonal entries in J_M^* are bounded by λ^*. Therefore according to the definition of $\lambda_\delta := \lambda - \lambda^*$, the entries in $\big(\widetilde{J}_M\big)_{\text{off},S}$ are bounded by λ and therefore the condition (a) for feasibility of primal-dual witness method is satisfied, i.e., we have $\|\widetilde{J}_M\|_{\infty,\text{off},S} \le \lambda$. Second, since $\|\widetilde{\Delta}_J\|_{\infty,S_R} = \lambda_\delta$, we have $\|\widetilde{\Delta}_J\|_{\infty,S} \le \|\widetilde{\Delta}_J\|_{\infty,S_R}$ and therefore $\|\widetilde{\Delta}_J\|_\infty = \|\widetilde{\Delta}_J\|_{\infty,S_R} = \lambda_\delta$ which results the following error bound

$$
\|\widetilde{\Delta}_J\|_\infty := \|\widetilde{J}_M - J_M^*\|_\infty \le 4K_{SS}\Big(1+\frac{m}{\alpha}\Big)\overline{\delta}_f(p^\tau;n). \tag{34}
$$

Furthermore, $\|\widetilde{\Delta}_J\|_\infty < \frac{1}{ldK_M}$ bound can be concluded from above inequality by substituting $\overline{\delta}_f(p^\tau;n)$ from (32). Thus, the condition for Lemma 6 is satisfied and we have the following bound on the remainder term

$$
\begin{aligned}
\|R(\widetilde{\Delta}_J)\|_\infty &\le \frac{l}{l-1}d\|\widetilde{\Delta}_J\|_\infty^2 K_M^3\\
&\le \frac{16l}{l-1}dK_M^3K_{SS}^2\Big(1+\frac{m}{\alpha}\Big)^2\big[\overline{\delta}_f(p^\tau;n)\big]^2\\
&= \left[\frac{16l}{l-1}dK_M^3K_{SS}^2\Big(1+\frac{m}{\alpha}\Big)^2\overline{\delta}_f(p^\tau;n)\right]\overline{\delta}_f(p^\tau;n)\\
&\le \overline{\delta}_f(p^\tau;n) = \frac{\alpha}{m}\gamma,
\end{aligned}
$$

where in the second inequality, we used error bound in (34) and the last inequality is concluded from bound (32).

Now the conditions for Lemma 5 are satisfied and therefore we have the upper bound on $\|\widetilde{\Delta}_R\|_{\infty,S_R} < C_3\gamma$ and the strict dual feasibility on S_M^c. Second result satisfies condition (c) of the primal-dual witness method feasibility conditions. The upper bound on $\|\widetilde{\Delta}_R\|_{\infty,S_R}$ in

conjunction with the lower bound on $\left(\Sigma_R^*\right)_{\min} > C_3\gamma$ (mentioned in the theorem), ensures that the sign of Σ_R^* and $\widetilde{\Sigma}_R$ are the same which results that the condition (b) of the feasibility conditions for primal-dual witness method is satisfied. Since all three conditions (a)-(c) are satisfied, we have equivalence between the modified program and the original one under conditions specified in the theorem. It gives us both results (a) and (b) in the theorem. Then by assuming lower bound on minimum nonzero value of J_M^*, the result in part (c) is also proved.

As mentioned before, we need to show that the dropped constraint $\Sigma = \Sigma_M - \Sigma_R \succ 0$ is also satisfied. Since the conditions for Corollary 13 in Appendix E.5 are satisfied, we have the spectral norm error bound (41) on overall covariance matrix Σ. Applying the inverse tail function for Gaussian distribution in (35) to assumption (A.6) results that the minimum eigenvalue of exact covariance matrix Σ^* satisfies lower bound $\lambda_{\min}(\Sigma^*) \geq \left(C_4 + \frac{m}{\alpha}C_3\right)d\overline{\delta}_f(p^\tau; n) + C_5 d^2\left[\overline{\delta}_f(p^\tau; n)\right]^2$ where $C_6 := \left(C_4 + \frac{m}{\alpha}C_3\right)\sqrt{2q^2}$ and $C_7 := 2q^2C_5$. Then by exploiting Weyl's theorem (Theorem 4.3.1 in Horn and Johnson (1985)), the estimated covariance matrix $\widehat{\Sigma}$ is positive definite and thus valid. Therefore, the result is proved. ■

Appendix E. Auxiliary Lemmata

First, the tail condition for a probability distribution is defined as follows.

Definition 12 (Tail Condition) *The random vector X satisfies tail condition with parameters f and v_* if there exists a constant $v_* \in (0, \infty)$ and function $f : \mathbb{N} \times (0, \infty) \to (0, \infty)$ such that for any $(i, j) \in V \times V$:*

$$\mathbb{P}[|\widehat{\Sigma}^n - \Sigma_{ij}^*| \geq \delta] \leq \frac{1}{f(n, \delta)} \text{ for all } \delta \in (0, \frac{1}{v_*}].$$

Note that since the function $f(n, \delta)$ is an increasing function of both variables n and δ, we define the inverse functions $\overline{n}_f(r; \delta)$ and $\overline{\delta}_f(r; n)$ with respect to variables n and δ respectively (when the other argument is fixed), where $f(n, \delta) = r$.

E.1 Concentration Bounds

From Lemma 1 in Ravikumar et al. (2011), we have the following concentration bound for the empirical covariance matrix of Gaussian random variables.

Lemma 3 (Ravikumar et al. 2011) *Consider a set of Gaussian random variables with covariance matrix Σ^*. Given n i.i.d. samples, the sample covariance matrix $\widehat{\Sigma}^n$ satisfies*

$$\mathbb{P}[|\widehat{\Sigma}_{ij}^n - \Sigma_{ij}^*| > \delta] \leq 4\exp\left\{-\frac{n\delta^2}{2q^2}\right\} \text{ for all } \delta \in (0, q),$$

for some constant $q > 0$.

Thus the tail function for Gaussian random vector takes the exponential form with the following corresponding inverse functions:

$$\overline{n}_f(r;\delta) = \frac{2q^2\log(4r)}{\delta^2}, \quad \overline{\delta}_f(r;n) = \sqrt{\frac{2q^2\log(4r)}{n}} \tag{35}$$

Applying above Lemma, we get the following bound for sampling error.

Lemma 4 (Ravikumar et al. 2011) *For any $\tau > 2$ and sample size n such that $\overline{\delta}_f(p^\tau;n) < 1/v_*$, we have*

$$\mathbb{P}\big[\|E\|_\infty \geq \overline{\delta}_f(p^\tau;n)\big] \leq \frac{1}{p^{\tau-2}} \to 0.$$

E.2 Feasibility Conditions

In the following lemma, we propose some conditions to bound the residual error $\|\widetilde{\Delta}_R\|_{\infty,S_R}$ and also satisfy the condition (c) of feasibility conditions required for equivalence between the witness solution and the original one.

Lemma 5 *Suppose that*

$$\begin{aligned} \max\{\|R\|_\infty, \|E\|_\infty\} &\leq \frac{\alpha}{m}\gamma, \\ \lambda_\delta &= \frac{2K_{SS}}{1-2K_{SS_R}}\Big(1+\frac{\alpha}{m}\Big)\gamma, \end{aligned} \tag{36}$$

then

a) $\|\widetilde{\Delta}_R\|_{\infty,S_R} \leq C_3\gamma$ *for some* $C_3 > 0$.

b) $\|\widetilde{Z}\|_{\infty,S_M^c} < \gamma$.

Proof Applying definitions (29) and (30) to optimality condition considered in second step of primal-dual witness method construction, gives the following equivalent equation

$$J_M^{*-1}\widetilde{\Delta}_J J_M^{*-1} - \Sigma_R^* - R\big(\widetilde{\Delta}_J\big) + E + \widetilde{Z} = 0. \tag{37}$$

Above equation is a $p \times p$ matrix equation. We can rewrite it as a linear equation with size p^2 if we use the vectorized form of matrices. Vectorized form of a matrix $D \in \mathbb{R}^{p\times p}$ is a column vector $\overline{D} \in \mathbb{R}^{p^2}$ which is composed by concatenating the rows of matrix D in a single column vector. In the vectorized form, we have

$$\mathrm{vec}\big(J_M^{*-1}\widetilde{\Delta}_J J_M^{*-1}\big) = \big(J_M^{*-1}\otimes J_M^{*-1}\big)\overline{\widetilde{\Delta}}_J = \Gamma^*\overline{\widetilde{\Delta}}_J.$$

Decomposing the vectorized form of (37) into three disjoint partitions S, S_R and S_M^c gives the following decomposed form

$$\begin{bmatrix} \Gamma^*_{SS} & \Gamma^*_{SS_R} & \Gamma^*_{SS_M^c} \\ \Gamma^*_{S_RS} & \Gamma^*_{S_RS_R} & \Gamma^*_{S_RS_M^c} \\ \Gamma^*_{S_M^cS} & \Gamma^*_{S_M^cS_R} & \Gamma^*_{S_M^cS_M^c} \end{bmatrix} \begin{bmatrix} \big(\overline{\widetilde{\Delta}}_J\big)_S \\ \overrightarrow{\lambda_\delta} \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ \big(\overline{\Sigma}^*_R\big)_{S_R} \\ 0 \end{bmatrix} + \begin{bmatrix} \big(-\overline{R}+\overline{E}+\overline{\widetilde{Z}}\big)_S \\ \big(-\overline{R}+\overline{E}+\overline{\widetilde{Z}}\big)_{S_R} \\ \big(-\overline{R}+\overline{E}+\overline{\widetilde{Z}}\big)_{S_M^c} \end{bmatrix} = 0, \tag{38}$$

where we used the equalities $\big(\overline{\widetilde{\Delta}}_J\big)_{S_R} = \overrightarrow{\lambda_\delta}$ and $\big(\overline{\widetilde{\Delta}}_J\big)_{S_M^c} = 0$. Note that vector $\overrightarrow{\lambda_\delta}$ only includes $\pm\lambda_\delta$ entries according to the constraints in the modified program. Also note that Σ^*_R is zero in sets S and S_M^c. We also dropped the argument $\widetilde{\Delta}_J$ from remainder function $R\big(\widetilde{\Delta}_J\big)$ to simplify the notation.

Similar to the original program, the matrix $\widetilde{Z}$ is composed of two parts, $\widetilde{Z}_\beta$ and $\widetilde{Z}_\gamma$, i.e., $\widetilde{Z} = \widetilde{Z}_\beta + \gamma\widetilde{Z}_\gamma$. Matrix $\widetilde{Z}_\beta = \widetilde{\Sigma}_R$ from equation (22), includes Lagrangian multipliers and $\widetilde{Z}_\gamma \in \partial\|\widetilde{J}_M\|_{1,\text{off}}$. For set S, $\big(\widetilde{Z}_\beta\big)_S = 0$, since we don't have any constraint in the program and therefore the Lagrangian multipliers are zero. Applying this to the first row of equation (38) and since Γ^*_{SS} is invertible, we have the following for error $\overline{\widetilde{\Delta}}_J$ in set S

$$\big(\overline{\widetilde{\Delta}}_J\big)_S = {\Gamma^*_{SS}}^{-1}\Big[-\Gamma^*_{SS_R}\overrightarrow{\lambda_\delta} + \overline{R}_S - \overline{E}_S - \gamma\big(\overline{\widetilde{Z}}_\gamma\big)_S\Big], \tag{39}$$

In set S_R, $\overline{\widetilde{Z}}_{S_R} = \big(\overline{\widetilde{\Sigma}}_R\big)_{S_R} + \gamma\big(\overline{\widetilde{Z}}_\gamma\big)_{S_R}$. Applying this to the second row of equation (38) results

$$\Gamma^*_{S_RS}\big(\overline{\widetilde{\Delta}}_J\big)_S + \Gamma^*_{S_RS_R}\overrightarrow{\lambda_\delta} + \big(\overline{\widetilde{\Delta}}_R\big)_{S_R} + \gamma\big(\overline{\widetilde{Z}}_\gamma\big)_{S_R} - \overline{R}_{S_R} + \overline{E}_{S_R} = 0,$$

Recall that we defined $\widetilde{\Delta}_R := \widetilde{\Sigma}_R - \Sigma^*_R$. Substituting (39) in above equation results the following for error $\overline{\widetilde{\Delta}}_R$ in set S_R

$$\begin{aligned}\big(\overline{\widetilde{\Delta}}_R\big)_{S_R} =& -\Gamma^*_{S_RS}{\Gamma^*_{SS}}^{-1}\Big[-\Gamma^*_{SS_R}\overrightarrow{\lambda_\delta} + \overline{R}_S - \overline{E}_S - \gamma\big(\overline{\widetilde{Z}}_\gamma\big)_S\Big]\\ &- \Gamma^*_{S_RS_R}\overrightarrow{\lambda_\delta} - \gamma\big(\overline{\widetilde{Z}}_\gamma\big)_{S_R} + \overline{R}_{S_R} - \overline{E}_{S_R}.\end{aligned}$$

Taking ℓ_∞ element-wise norm from above equation and using inequality $\|Ax\|_\infty \le |\!|\!|A|\!|\!|_\infty\|x\|_\infty$ for any matrix $A \in \mathbb{R}^{r\times s}$ and vector $x \in \mathbb{R}^s$, results the bound

$$\begin{aligned}\|\widetilde{\Delta}_R\|_{\infty,S_R} \le |\!|\!|-\Gamma^*_{S_RS}{\Gamma^*_{SS}}^{-1}\Gamma^*_{SS_R} + \Gamma^*_{S_RS_R}|\!|\!|_\infty\lambda_\delta + |\!|\!|\Gamma^*_{S_RS}{\Gamma^*_{SS}}^{-1}|\!|\!|_\infty\big[\|\overline{R}_S\|_\infty + \|\overline{E}_S\|_\infty + \gamma\big]\\ + \big(\|\overline{R}_{S_R}\|_\infty + \|\overline{E}_{S_R}\|_\infty + \gamma\big),\end{aligned}$$

where we used the fact that $\|\overrightarrow{\lambda_\delta}\|_\infty = \lambda_\delta$ and $\|\widetilde{Z}_\gamma\|_\infty = 1$. Now if we apply the assumptions mentioned in the lemma,

$$\begin{aligned}\|\widetilde{\Delta}_R\|_{\infty,S_R} \le \Bigg[\frac{2K_{SS}(m+\alpha)}{m(1-2K_{SS_R})}|\!|\!|-\Gamma^*_{S_RS}{\Gamma^*_{SS}}^{-1}\Gamma^*_{SS_R} + \Gamma^*_{S_RS_R}|\!|\!|_\infty\\ + \big(1+\frac{2\alpha}{m}\big)\big(1+|\!|\!|\Gamma^*_{S_RS}{\Gamma^*_{SS}}^{-1}|\!|\!|_\infty\big)\Bigg]\gamma = C_3\gamma,\end{aligned} \tag{40}$$

which proves part (a) of the Lemma.

Now if we substitute (39) in the equation from third row of (38), we have

$$\overline{\widetilde{Z}}_{S_M^c} = -\Gamma^*_{S_M^cS}{\Gamma^*_{SS}}^{-1}\Big[-\Gamma^*_{SS_R}\overrightarrow{\lambda_\delta} + \overline{R}_S - \overline{E}_S - \gamma\big(\overline{\widetilde{Z}}_\gamma\big)_S\Big] - \Gamma^*_{S_M^cS_R}\overrightarrow{\lambda_\delta} + \overline{R}_{S_M^c} - \overline{E}_{S_M^c}.$$

Taking ℓ_∞ element-wise norm from above equation gives the following bound

$$\|\widetilde{Z}\|_{\infty,S_M^c} \le |\!|\!|\Gamma^*_{S_M^cS}{\Gamma^*_{SS}}^{-1}\Gamma^*_{SS_R} - \Gamma^*_{S_M^cS_R}|\!|\!|_\infty\lambda_\delta + |\!|\!|\Gamma^*_{S_M^cS}{\Gamma^*_{SS}}^{-1}|\!|\!|_\infty\big[\|\overline{R}_S\|_\infty + \|\overline{E}_S\|_\infty + \gamma\big]$$

$$+ \|\overline{R}_{S_M^c}\|_\infty + \|\overline{E}_{S_M^c}\|_\infty,$$

where we used the fact that $\|\widetilde{Z}_\gamma\|_\infty = 1$. Applying assumption (A.4) to above bound results

$$\|\widetilde{Z}\|_{\infty,S_M^c} \leq (1-\alpha)\lambda_\delta + (2-\alpha)\left[\|R\|_\infty + \|E\|_\infty\right] + (1-\alpha)\gamma.$$

Using assumptions stated in the Lemma, we have

$$\begin{aligned}
\|\widetilde{Z}\|_{\infty,S_M^c} &\leq \left[\frac{2K_{SS}}{1-2K_{SS_R}}\Big(1+\frac{\alpha}{m}\Big)(1-\alpha) + (2-\alpha)\frac{2\alpha}{m} + (1-\alpha)\right]\gamma \\
&< \left[4K_{SS}\Big(1+\frac{\alpha}{m}\Big)(1-\alpha) + (2-\alpha)\frac{2\alpha}{m} + (1-\alpha)\right]\gamma \\
&< \left[4K_{SS}\frac{m-(m-1)\alpha}{m} + \frac{4\alpha}{m} + (1-\alpha)\right]\gamma \leq \gamma,
\end{aligned}$$

where we used the bound on K_{SS_R} in assumption (A.4) in the second inequality and the fact that $\alpha > 0$ in the third inequality. Final inequality is derived from assumption (A.5) which finishes the proof of part (b). ■

E.3 Control of Remainder

In the following Lemma which is stated and proved in lemma 5 in Ravikumar et al. (2011), the argument $\widetilde{\Delta}_J$ controls the remainder function behavior.

Lemma 6 *Suppose that the element-wise ℓ_∞ bound $\|\widetilde{\Delta}_J\|_\infty \leq \frac{1}{lK_M d}$ for some $l > 1$ holds. Then*

$$R\big(\widetilde{\Delta}_J\big) = \big({J_M^*}^{-1}\widetilde{\Delta}_J\big)^2 Q {J_M^*}^{-1},$$

where $Q := \sum_{k=0}^{\infty}(-1)^k\big({J_M^*}^{-1}\widetilde{\Delta}_J\big)^k$ with bound $\|\!|Q^T|\!\|_\infty \leq \frac{l}{l-1}$. Also, in terms of element-wise ℓ_∞ norm, we have

$$\|R\big(\widetilde{\Delta}_J\big)\|_\infty \leq \frac{l}{l-1} d\|\widetilde{\Delta}_J\|_\infty^2 K_M^3.$$

E.4 Control of $\widetilde{\Delta}_J$

According to the primal-dual witness solutions construction, we have the error bounds on $\widetilde{\Delta}_J$ within the sets S_R and S_M^c such that $\|\widetilde{\Delta}_J\|_{\infty,S_R} = \lambda_\delta$ and $\|\widetilde{\Delta}_J\|_{\infty,S_M^c} = 0$. In the following lemma, we propose some conditions to control the error $\|\widetilde{\Delta}_J\|_{\infty,S}$.

Lemma 7 *Suppose that*

$$r := 2K_{SS_R}\lambda_\delta + 2K_{SS}\big(\|E\|_\infty + \gamma\big) \leq \min\left\{\frac{1}{ldK_M}, \frac{l-1}{2ldK_{SS}K_M^3}\right\},$$

then we have the following element-wise ℓ_∞ bound for $\big(\widetilde{\Delta}_J\big)_S$,

$$\|\widetilde{\Delta}_J\|_{\infty,S} \leq r.$$

The proof is within the same lines of Lemma 6 proof in Ravikumar et al. (2011) but with some modifications since the error $\|\widetilde{\Delta}_J\|_{\infty,S_R}$ is not zero and therefore the nonzero value λ_δ arises in the final result. Since the modified optimization program (21) is different with the modified program in Ravikumar et al. (2011), it is worth discussing about existing a unique solution for the modified optimization program (21). This uniqueness can be shown with similar discussion presented in Appendix B for uniqueness of the solution of original program (19). We only need to show that there is no problem in uniqueness by removing the off-diagonal constaraints for set S in the modified program. By Lagrangian duality, the ℓ_1 penalty term $\gamma\|J_M\|_{1,\text{off}}$ can be moved to constraints as $\|J_M\|_{1,\text{off}} \leq C(\gamma)$ for some bounded $C(\gamma)$. Therefore, the off-diagonal entries in set S where the corresponding constraints were relaxed in the modified program are still bounded because of this ℓ_1 constraint. Hence, the modified program (21) has a unique solution.

E.5 Spectral Norm Error Bound on Overall Covariance Matrix $\Sigma = J_M^{-1} - \Sigma_R$

Corollary 13 *Under the same assumptions (excluding (A.6)) as Theorem 5, with probability greater than $1 - 1/p^c$, the overall covariance matrix estimate $\widehat{\Sigma} = \widehat{\Sigma}_M - \widehat{\Sigma}_R$ satisfies spectral norm error bound*

$$\|\!|\widehat{\Sigma} - \Sigma^*|\!\| \leq \Big(C_4 + \frac{m}{\alpha}C_3\Big)d\overline{\delta}_f(p^\tau;n) + C_5 d^2\big[\overline{\delta}_f(p^\tau;n)\big]^2. \tag{41}$$

Proof We first bound the spectral norm errors for the Markov and residual covariance matrices $\widehat{\Sigma}_M$ and $\widehat{\Sigma}_R$. Along the same lines as Corollary 4 proof in Ravikumar et al. (2011), the spectral norm error $\|\!|\widehat{\Sigma}_M - \Sigma_M^*|\!\|$ can be bounded as

$$\|\!|\widehat{\Sigma}_M - \Sigma_M^*|\!\| \leq C_4 d\overline{\delta}_f(p^\tau;n) + C_5 d^2\big[\overline{\delta}_f(p^\tau;n)\big]^2,$$

where $C_4 = 4\big(1+\frac{m}{\alpha}\big)K_{SS}K_M^2$ and $C_5 = \frac{16l}{l-1}\big(1+\frac{m}{\alpha}\big)^2 K_{SS}^2 K_M^3$.
The spectral norm error $\|\!|\widehat{\Sigma}_R - \Sigma_R^*|\!\|$ can be also bounded as

$$\|\!|\widehat{\Sigma}_R - \Sigma_R^*|\!\| \leq \|\!|\widehat{\Sigma}_R - \Sigma_R^*|\!\|_\infty \leq d\|\widehat{\Sigma}_R - \Sigma_R^*\|_\infty \leq \frac{m}{\alpha}C_3 d\overline{\delta}_f(p^\tau;n),$$

where the first inequality is the property of spectral norm which is bounded by ℓ_∞-operator norm, second inequality is a result of the fact that $\widehat{\Sigma}_R$ and Σ_R^* has at most d nonzero entries in each row (since $S_R \subset S_M$) and the last inequality is concluded from the upper bound on ℓ_∞ element-wise norm error on residual matrix estimation stated in part (a) of Theorem 5. Applying the above bounds to the overall covariance matrix estimation $\widehat{\Sigma} = \widehat{\Sigma}_M - \widehat{\Sigma}_R$ and using the triangular inequality for norms, the bound in (41) is proven. ■

Appendix F. Proof of Corollary 9

Proof The result in this corollary is a special case of general result in Theorem 5 when $\lambda^* = 0$ and some minor modifications are considered in problem formulation. Note that, it is expressed in assumption (A.1) that the off-diagonal entries of exact Markov matrix J_M^* are upper bounded by some positive λ^*. In order to extend the proof to the case of $\lambda^* = 0$ (The

case in this corollary), we need some minor modifications. First, the identifiability assumptions (A.0)-(A.3) can be ignored and instead it is assumed that the Markov part J_M^* (or equivalently Σ_M^*) is diagonal and the residual part Σ_R^* has only nonzero off-diagonal entries. Since the diagonal Markov matrix and off-diagonal residual matrix do not have any nonzero overlapping entries, it is natural that we do not require any more identifiability assumptions. Then, with these new assumptions, the set S_M is defined as $S_M := S_R \cup \{(i,i)|i = 1, ..., p\}$ where S_R is defined the same as (10) and also set S is defined the same as (11) which results that set S includes only diagonal entries. Thus, the off-diagonal entries belongs to sets S_R and S_M^c. Since Σ_M^* is a diagonal matrix, all submatrices of Γ^* which are indexed by sets S_R or S_M^c are complete zero matrices. The result is that the terms which are bounded in the mutual incoherence condition (A.4) are already zero and thus there is no need to consider those additional assumptions in the corollary.
By making these changes in the problem formulation, the result in Corollary 9 can be proven within the same lines of general result proof in Theorem 5. It is only required to change the constraint on set S_R in the modified optimization program to $\big(J_M\big)_{S_R} = \lambda \operatorname{sign}\Big(\big(\Sigma_R^*\big)_{S_R}\Big)$. ■

References

A. Anandkumar, V. Y. F. Tan, and A. S. Willsky. High-dimensional Gaussian graphical model selection: Tractable graph families. *Preprint, ArXiv 1107.1270*, June 2011.

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc., New York, NY, 1984.

M. Banerjee and T. Richardson. On a dualization of graphical Gaussian models: A correction note. *Scandinavian Journal of Statistics*, 30(4):817–820, 2003.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. of Machine Learning Research*, 9:485–516, March 2008.

P.J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

S.P. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

L.D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. *Lecture Notes-Monograph Series, Institute of Mathematical Statistics*, 9, 1986.

T.T. Cai, C.H. Zhang, and H.H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *CoRR*, abs/0912.3599, 2009.

V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky. Rank-sparsity incoherence for matrix decomposition. *Arxiv preprint arXiv:0906.2220*, 2009.

V. Chandrasekaran, P.A. Parrilo, and A.S. Willsky. Latent variable graphical model selection via convex optimization. *Arxiv preprint arXiv:1008.1290*, 2010a.

V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Arxiv preprint arXiv:1012.0621*, 2010b.

M.J. Choi, V. Chandrasekaran, and A.S. Willsky. Gaussian multiresolution models: Exploiting sparse Markov and covariance structure. *Signal Processing, IEEE Transactions on*, 58(3):1012–1024, 2010.

M.J. Choi, V.Y.F. Tan, A. Anandkumar, and A. Willsky. Learning latent tree graphical models. *J. of Machine Learning Research*, 12:1771–1812, May 2011.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

D.R. Cox and N. Wermuth. Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218, 1993.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM. J. Matrix Anal. & Appl.*, 30(56), 2008.

A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2007.

R.A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, NY, 1985.

J.Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1), 2006.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. of Machine Learning Research*, 8:613–636, 2007.

G. Kauermann. On a dualization of graphical Gaussian models. *Scandinavian journal of statistics*, pages 105–116, 1996.

M. Kolar, A.P. Parikh, and E.P. Xing. On sparse nonparametric conditional covariance selection. In *International Conference on Machine Learning*, 2010.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

S.L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. of Machine Learning Research*, 10:2295–2328, 2009.

P.L. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Neural Information Processing*, 2011.

D.M. Malioutov, J.K. Johnson, and A.S. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *J. of Machine Learning Research*, 7:2031–2064, 2006.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

K. Mohan. ADMM algorithm for graphical lasso with an ℓ_∞ element-wise norm constraint. *arXiv:1311.7198*, Nov. 2013.

S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Arxiv preprint arXiv:1010.2731*, 2010.

P. Ravikumar, M.J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, (4):935–980, 2011.

A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

M.K. Silva, N.J.A. Harvey, and C.M. Sato. Sparse sums of positive semidefinite matrices. *Arxiv preprint arXiv:1107.0088*, 2011.

P. Spirtes and C. Meek. Learning Bayesian networks with discrete variables from data. In *Proc. of Intl. Conf. on Knowledge Discovery and Data Mining*, pages 294–299, 1995.

AS Wagaman and E. Levina. Discovering sparse covariance structures with the isomap. *J. of Computational and Graphical Statistics*, 18(3):551–572, 2009.

M.J. Wainwright and M.I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Proc. of NIPS*, 2008.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *J. of Machine Learning Research*, 10:555–568, 2009.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.